Machine Learning: Science and Technology

PAPER • OPEN ACCESS

Uncertainty quantification by ensemble learning for computational optical form measurements

To cite this article: Lara Hoffmann *et al* 2021 *Mach. Learn.: Sci. Technol.* **2** 035030

MACHINE
LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
5 February 2021

REVISED
5 May 2021

ACCEPTED FOR PUBLICATION
24 May 2021

PUBLISHED
13 July 2021

PAPER

Uncertainty quantification by ensemble learning for computational optical form measurements

Lara Hoffmann[*], **Ines Fortmeier and Clemens Elster**

Physikalisch-Technische Bundesanstalt, Braunschweig and Berlin, Germany

[*] Author to whom any correspondence should be addressed.

E-mail: lara.hoffmann@ptb.de

Keywords: deep learning, ensemble methods, uncertainty quantification, computational optical form measurement

Abstract

Uncertainty quantification by ensemble learning is explored in terms of an application known from the field of computational optical form measurements. The application requires solving a large-scale, nonlinear inverse problem. Ensemble learning is used to extend the scope of a recently developed deep learning approach for this problem in order to provide an uncertainty quantification of the solution to the inverse problem predicted by the deep learning method. By systematically inserting out-of-distribution errors as well as noisy data, the reliability of the developed uncertainty quantification is explored. Results are encouraging and the proposed application exemplifies the ability of ensemble methods to make trustworthy predictions on the basis of high-dimensional data in a real-world context.

1. Introduction

Artificial intelligence has established a major impact on science and applications. In particular, deep neural networks [39] show a great potential of understanding complex scientific relationships due to their deep and nonlinear structure. They have been successfully applied to various tasks including natural language processing [52], computational imaging [8] and data mining [49]. Deep learning applications in the domain of physics and engineering include solving partial differential equations [34], phase imaging [11] and misalignment calibration [53].

However, their ‘black-box’ character and the resulting lack of trustworthiness are probably the most crucial shortcomings of deep learning approaches. Many examples exist that demonstrate unreasonable behavior of trained networks. For instance, the technique of layer-wise relevance propagation has revealed that an apparently well-trained image classification network had adopted a ‘Clever Hans’ decision strategy [31]; the network had learned to classify horse images correctly by focusing on the bottom left corner of the image—there was an unnoticed tag remaining on horse images in the training data set. Adversarial attacks can mislead trained networks into making unreasonable predictions by only slightly perturbing the input data [2]. In 2018, an autonomous driving car crashed into a pedestrian in Arizona because the self-driving system did not classify her correctly [37].

These examples demonstrate the importance of understanding the behavior of deep neural networks in order to ensure their trustworthiness. Much effort has been devoted to developing corresponding approaches. Adadi and Berrada [1], Lapuschkin *et al* [31] and Selvaraju *et al* [45] propose and analyze different methods to explain the behavior of a network, making its predictions more transparent and easier to interpret. In [35], the Fisher information was used to detect unusual input to the network. Intense testing is another way of analyzing the behavior of a network trained on critical data and exploring its generalization capacity, cf [46, 47, 50].

Uncertainty quantification is also an important pillar to improve the trustworthiness of predictions made by a trained network [28]. Various approaches exist and there are different kinds of uncertainties to consider. Uncertainties are often classified as epistemic and aleatoric [18, 23]. Sources of uncertainty include imperfect training, unexpected shifts in the data, systematic errors and out-of-distribution data, to mention just a few (cf [3, 25, 38, 41]). The most common approaches to uncertainty quantification include Bayesian neural networks [29, 51], dropout based methods [18, 27] and ensemble techniques [12, 30, 33]. We focus on the latter, because ensemble learning scales well to higher dimensional data, performs best in recent uncertainty studies and is straightforward to implement [10, 20, 38, 41].

The goal of this paper is to explore the potential of ensemble techniques for uncertainty quantification in deep learning in terms of a large-scale inverse problem known from computational optical form measurements. The considered application is based on the tilted-wave interferometer (TWI), which is an accurate, interferometric measurement system for the form measurement of optical aspheres and freeform surfaces [5, 17]. Applied conventional methods solve the high-dimensional, nonlinear, inverse reconstruction problem iteratively through local linearizations. In general, uncertainty quantification is a growing field in computational science and engineering. Considering the TWI application, there exists previous work in which quantities have been identified that influence the uncertainty [15]. Some of these quantities have been analyzed via experimental measurements, Monte-Carlo simulations or the Jacobian used for the iterative topography updates [6, 16, 17, 43]. There has also been work to increase the robustness of the TWI evaluation procedure by removing nonrotationally symmetric errors [44]. A recent study comparing different state-of-the-art techniques that measure optical freeform surfaces [14] showed deviations between the participants ranging from 15 to 110 nm on some real specimens neglecting any spherical contribution.

The novelty of this paper is twofold. First, we extend a previous deep learning approach for this application [22] to incorporate an uncertainty quantification of its predictions. This is achieved through ensemble learning. In contrast to [22], the networks are trained on a calibrated data set, which will be discussed in more detail. Second, we systematically insert an increasing out-of-distribution calibration error into the system and analyze its effect on the reliability of the developed uncertainty quantification. Furthermore, the influence of noise is investigated. So far, high-dimensional uncertainty quantification for scalable deep learning techniques is hardly treated in literature [20], which makes the results of our chosen application useful for other machine learning applications as well.

The paper is structured as follows. Section 2 introduces the chosen application from computational optical form measurements, followed by a detailed explanation of the data generation process in section 3. The employed deep neural network, ensemble learning and the corresponding uncertainty quantification are introduced in section 4. Results are then presented in section 5 with a particular focus on the impact of systematic calibration errors on the uncertainty quantification. Finally, in section 6, the potential benefit of deep learning for computational optical form measurements is discussed and possible future research motivated.

2. Application

The application of computational optical form measurements considered here focuses on measuring optical aspheres and freeform surfaces. This application is based on the [5], which will be introduced in the following.

The realization of the TWI considered here is the one employed at Physikalisch-Technische Bundesanstalt (PTB) [15, 17]. Its experimental set-up is shown in figure 1. The coherent light of a laser source (not shown in figure 1) is split into a reference arm and a measurement arm. In the measurement arm, the collimated light passes a 2D micro lens array. Each of these micro lenses acts like a point source, thus generating differently tilted wave fronts. After passing through the objective, the wavefronts are reflected at the surface under test and interfere at the beam splitter with the light from the reference arm. The resulting intensity images are captured on the charge-coupled device (CCD) and are unwrapped to optical path length differences using the Goldstein unwrapping algorithm [19]. A beam stop in the Fourier plane of the imaging optics prevents subsampling effects. Depending on the local slope of the specimen, different light sources generate resolvable sub-interferograms (patches) at the CCD. Information overlap at the CCD is prevented by using four disjoint masks on the point source array; this process eventually results in four images of optical path length differences for one specimen topography.

The goal is to measure the deviation of the specimen from its known design topography given the optical path length differences computed from the observed CCD intensities. The toolbox SimOptDevice [42] is used to model these optical path length differences in dependence on the topography of the specimen under test. The nonlinear inverse problem consists in finding the topography of the specimen such that the modeled optical path length differences best fit the observed ones.

The computer model of the optical system used to solve the inverse problem is not perfect and is usually ‘calibrated’ (i.e. adjusted) using the observed data (i.e. the optical path length differences) for some test specimens whose topographies are known with high accuracy. The calibration is realized by adding two virtual planes called *reference planes* to the computer model of the optical system, as shown in figure 2. The

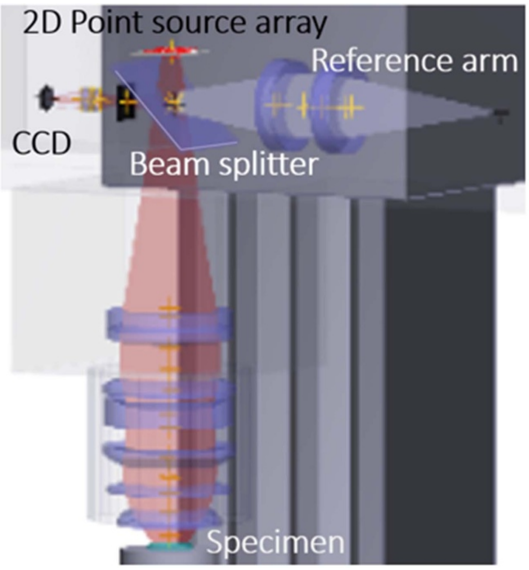

Figure 1. The experimental set-up of the TWI.

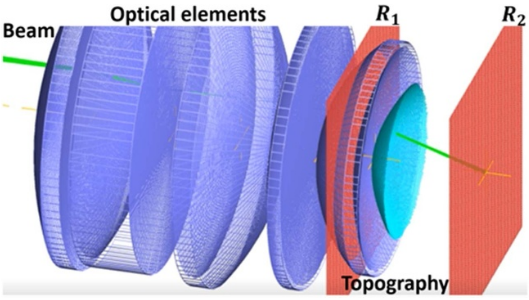

Figure 2. Two reference planes, R_1 and R_2, are added next to the topography in the computer modeled optical system for calibration.

light beam passes the first plane R_1 before attaining the topography and passes the second plane R_2 after having attained the topography. Each virtual reference plane modifies the light beams phenomenologically; the reference planes are parameterized by Zernike polynomials. Zernike polynomials are orthogonal on the unit disc and are commonly used in optics to represent wavefronts [48]. The coefficients of the Zernike polynomials are determined in such a way that the computer model best fits the optical path length differences deduced from the CCD intensities measured by the optical system for a chosen test specimen that is known with high accuracy [4]. Further details are given in appendix A.

3. Data generation

The goal of the TWI is to measure the deviation of any given specimen relative to its known design. The asphere we use as a design topography is characterized in appendix B. A data set containing various topographies is generated through randomly drawn sets of Zernike coefficients [48]. The Zernike polynomials parameterize the difference topographies ΔT, i.e. the deviation of the specimen from the known design. Then, the optical path length differences through the optical system are computed for the design topography and each generated specimen at a fixed position, respectively. This is realized with the simulation toolbox SimOptDevice [42]. Hence, for each specimen, a difference of optical path length differences ΔL is obtained.

Each sample in the generated data set consists of a set of differences of optical path length differences ΔL and the difference topography ΔT between the specimen and the design topography. An example is shown in figure 3. The difference of optical path length differences consists of four images because four disjoint masks are sequentially used on the 2D point source array. In total, almost 40 000 (virtual) topographies are generated for training and about 2000 are generated for testing. The mean root mean squared deviation from

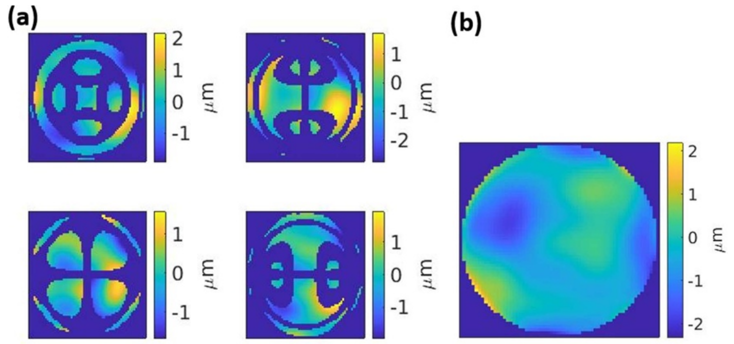

Figure 3. The difference of optical path length differences ΔL is shown in (a) and the corresponding difference topography ΔT is shown in (b). The aperture of the specimen has a radius of about 15 mm.

the design topography of the generated topographies in the test data set is 564 nm and their median deviation is 473 nm. The peak to valley differences of the topographies of the test set range from 42 nm to 11.6 μm. Additional examples are shown in appendix C (figure 16) to illustrate the diversity of the test data. The exact amount of training and test data are not optimized. It was shown in [22] (figure 8) that the number of available training data is crucial for the prediction capability. We found that 40 000 training data are enough to achieve accurate predictions. More data would likely lead to even better results, but the goal of this paper is to systematically analyze the uncertainty quantification of ensemble learning. We also found that the generated test set already represents a large diversity of possible difference topographies.

In [22], the training data were generated without including reference planes to the model of the optical system, and the simulated data were constructed under the assumption of a perfect model for the optical system. In this paper, systematic investigations on the impact of calibration errors are carried out. For this purpose, the test data are generated using a non-perfect optical system by deliberately adding calibration errors to the optical system. These errors represent errors caused by an imperfect calibration and will be termed *calibration errors* in the following. To construct test data containing such calibration errors, the virtual reference planes of the perfectly calibrated optical system are systematically modified, which alters the beam path through the optical system, cf figure 2. Recall that the training data were determined by simulating data for the perfectly calibrated optical system.

Investigating the predictive ability of the trained net for test data corrupted by calibration errors allows the generalizability of the trained net to be explored for a more realistic scenario. From the point of view of machine learning, the test set constructed under calibration errors is an out-of-distribution test set, as no calibration errors are contained in the training set. We are particularly interested in the behavior of the calculated uncertainties and in the extent to which they reflect the errors in the reconstructed topographies caused by the calibration errors. In total, the introduced calibration error affects the differences of optical path length differences ΔL up to a root mean squared deviation of 219 nm on the test data set. We refer to appendix A for further details about the construction of the test data.

Noisy test data are generated by adding Gaussian noise to the input of the existing test data. Note that the training set is fixed and does not adapt to errors introduced into the test set.

4. Method

The inverse problem at hand can be stated as follows. Find a map f, such that the difference of optical path length differences ΔL maps to the corresponding difference topography ΔT (see figure 3), i.e.:

$$f\colon \mathbb{R}^{4\times D\times D} \to \mathbb{R}^{D\times D},\ \Delta L \mapsto \Delta T, \tag{1}$$

where D is the given or chosen resolution of the images. Here, we choose D equal to 64. The choice in the dimensionality of the optical path length differences and difference topographies is not mandatory and done for convenience here. The function f can be approximated by a parameterization f_ϕ with parameter space Φ, solving the following minimization problem for all possible tuples $(\Delta L, \Delta T)$:

$$\min_{\phi\in\Phi} \|f_\phi(\Delta L) - \Delta T\|_2. \tag{2}$$

Recall that the difference topography ΔT is the difference between the known design topography and the specimen at hand. Equally, the difference of optical path length differences ΔL is the difference between, on

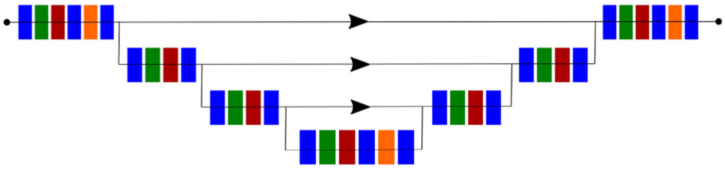

Figure 4. Example U-Net architecture.

the one hand, the optical path length difference (i.e. the difference between the optical path length of the 'measurement arm' and the 'reference arm') derived from the design topography and, on the other hand, the one measured on the specimen. In a real TWI measurement, the optical path length differences of the specimen are measured (based on the observed CCD values), while the optical path length differences of the design topography are calculated using the computer model of the optical system. For reasons of simplification, ΔT and ΔL will from now on be simply called topography and optical path length difference, respectively, as there is a unique dependency.

Neural networks have become a popular method used in imaging since the introduction of convolutional neural networks [32]. They can extract location invariant features and share weights which means that training requires fewer parameters. Deep neural networks, which are neural networks with many hidden layers, are commonly used today thanks to convolutional layers and technological advancement. The U-Net [40] is a specific deep neural network architecture which achieved accurate results in various imaging tasks (cf [8, 13] or [24]). Therefore, we chose the U-Net as a network architecture similar to [22]. An example of a U-Net structure is shown in figure 4. The input is processed through several layers, most of which are convolution layers, and results in the predicted output from left to right. The network has a bottleneck structure, which means that the dimension of the image data is reduced after each bundle of layers until it attains its minimal dimension in the center and increases afterwards in the same way. Furthermore, the arrows indicate skip connections between the bundles of layers of the same dimension, which means that the output of the last layer with the same dimension is depth concatenated to the input after the dimensional increase.

The parameterization f_ϕ is in general defined through the chosen neural network architecture and its concrete structure as for example the used activation functions, the actual depth of the network or the individual layers. All trainable network parameters are real numbers. We chose the ReLU activation function [36] for the U-Net because it is state-of-the-art for deep neural networks. An alternative parameterization could be one that predicts Zernike coefficients instead of the direct height values of the topography. However, here we are more interested in a direct image-to-image regression task and do not want to limit the prediction to a certain polynomial degree.

It is well known that trained neural networks typically find only a local and not a global optimum. Training multiple networks and making the prediction a decision of the constructed network ensemble [21] is a straightforward solution to overcome this problem to a certain extent. Deep ensembles have been proposed in [30] for uncertainty quantification because of their predictive variety. The random initialization, along with the random shuffling of data points during training, is considered to induce sufficient diversity to the network ensemble. We propose that dropout layers [7] also be included during training to encourage diversity. In contrast to [30], we focus on model uncertainty and do not predict an extra variance per output neuron, which would lead to many more learnables because of the high-dimensional output.

In our experience a lot of effort goes into finding an appropriate network architecture, especially when considering high-dimensional regression tasks. For example, simpler convolutional neural networks or autoencoders without skip-connections did not achieve accurate results in this application in contrast to the U-Net. A major advantage of network ensembles compared to other deep learning uncertainty methods is their scalability to high-dimensional data and the simplicity of implementation if a well suited architecture can be found. This is because the training procedure just has to be carried out multiple times without further adjustments. Other uncertainty methods result in a computationally more expensive approximate Bayesian inference [30] or require a modified loss function [18]. We use an ensemble of U-Nets trained with the mean squared error loss since we found that the mean squared error loss is more robust than for example the negative log likelihood loss function.

Previous work indicates that few ensemble members suffice to attain good results [3] and training time gets expensive with an increasing number of networks. We chose the ensemble to contain eight networks. This number is probably not optimal, but the prediction capability and uncertainty estimation of the ensemble seem to converge already as shown in table 2 in appendix C. The U-Nets are independently trained for the ensemble using the mean squared error as a loss function with the Adam optimizer [26] and an L^2 regularization with regularization factor 0.002. Note that by minimizing the chosen loss function, i.e. $\frac{1}{N'}\sum_{i=1}^{N'}\|f_{\phi'}(\Delta L_i)-\Delta T_i\|_2^2+\lambda\|\phi'\|_2^2$, where λ is the regularization parameter, N' is the current batch size and ϕ' the current network parameters, (2) is implicitly optimized. Each network consists of 69 layers, including five max pooling layers and transposed convolution layers, respectively. The initial learning rate equals 5×10^{-5}, with a learning rate drop factor of 0.75 every fourth epoch. In total, each network trains for 25 epochs with a mini batch size of 64, and the data are randomly shuffled every epoch. These parameters were selected after an initial grid search and tests with other choices of parameters. Similarly to the exact number of ensemble members and amount of training data, the selected parameters could be further optimized. However, we did not do so as we found that they serve our purpose sufficiently well. Training takes around 45 min for one network on a single GPU (Tesla V100-NVLINK on an Intel Xeon Gold 6226).

To sum up, we train an ensemble of $M=8$ deep neural networks $(f_{\phi_i})_i$, $\phi_i\in\Phi$, $i=1,\ldots,M$, that all have the same U-Net architecture and are independently trained on the same training data. Diversity is induced through random initialization, random data shuffling per epoch and inclusion of dropout layers during the training procedure. The ensemble prediction is defined as the average over all predictions, i.e.:

$$f_\phi(\Delta L) := \frac{1}{M}\sum_{j=1}^{M} f_{\phi_j}(\Delta L) \;\in\mathbb{R}^{D\times D}. \tag{3}$$

We define the *ensemble uncertainty* as the standard deviation over the predictions:

$$uc(f_\phi(\Delta L)) := \left(\frac{1}{M}\sum_{j=1}^{M}\left(f_{\phi_j}(\Delta L)-f_\phi(\Delta L)\right)^{.2}\right)^{.\frac{1}{2}} \in\mathbb{R}^{D\times D}. \tag{4}$$

In (4), $\cdot^{2}$ and $\cdot^{\frac{1}{2}}$ indicate elementwise square and square root, respectively. This definition is in line with the uncertainty definition from [30] when omitting the aleatoric part. If the ensemble uncertainty is considered for an entire topography, we refer to the *topography uncertainty* defined as $\sqrt{\frac{1}{D'}\sum_{d=1}^{D'}uc(f_\phi(\Delta L))_d^2}\in\mathbb{R}$, where D' is the number of pixel coordinates of the predicted topography ΔT.

The trained network ensemble takes about 300 ms to predict 30 difference topographies (on an Intel Xeon Gold 6226, single GPU, Tesla V100-NVLINK). In [17], the TWI evaluation procedure takes about 15 s for one topography reconstruction considering 13 350 rays, and assuming five iterative reconstruction steps (on an Intel(R), Core(TM), i7-2600, single CPU).

5. Results

This section presents the results obtained when considering a test set constructed using a perfectly calibrated system and test data produced by an optical system containing calibration errors. In addition, results are shown when the input data from the test set are corrupted by noise. Recall that the training data are constructed by simulating a perfectly calibrated optical system. The test set constructed after introducing calibration errors into the physical model used to generate the data can be seen as an out-of-distribution test and explores the generalizability of the trained network. Our focus lies on analyzing the uncertainty quantification produced by the network ensemble. As mentioned above, the uncertainty is primarily an estimation of the epistemic uncertainty and not the aleatoric uncertainty [18], as there is no noise in the simulated training data and because of the high-dimensional output. Nonetheless, the behavior of the ensemble uncertainty is analyzed on out-of-distribution test data by means of the systematically introduced calibration errors, and also through analyzing test data whose input is corrupted by white noise.

5.1. Perfectly calibrated system

The performance of the trained network ensemble is evaluated on the test data set generated by the same optical system as the training data, i.e. a perfectly calibrated system. The average root mean squared error equals 77 nm. First results of the ensemble prediction are shown in figure 5. Three example topographies with greatly varying sizes are reconstructed and the difference between the ground truth and the prediction is given. The network ensemble has no difficulties reconstructing the different topographies. The main error occurs at the edge of the topographies. This is not surprising, since the input data is more error-prone at the topography edges. There exists almost no redundant information since the patches rarely overlap at the edges, and some rays even leave the optical system without returning to the CCD. It is well known that outliers have a high impact on the L^2 measure. Therefore, the median error (i.e. the median of the absolute

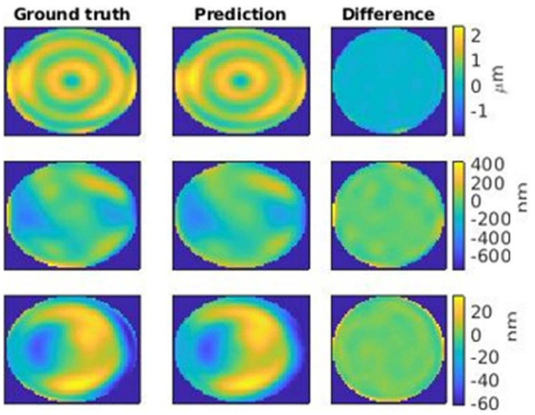

Figure 5. The ground truth, the ensemble prediction and their difference are shown for three example topographies with median errors of 44.5, 17.2 and 1.5 nm, respectively.

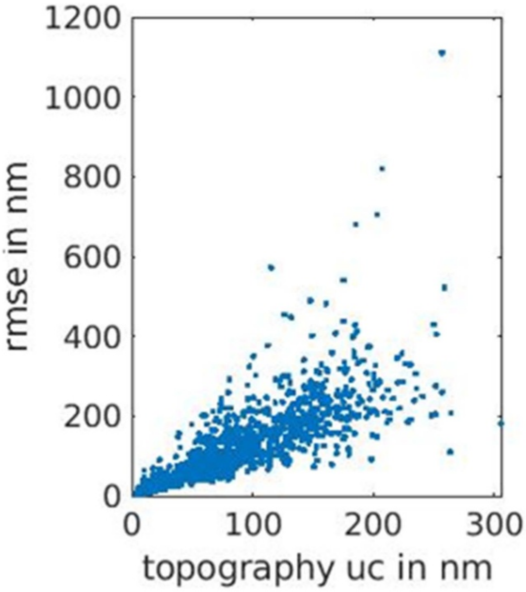

Figure 6. The root mean squared errors of the ensemble predictions of the topographies are plotted against their corresponding topography uncertainties.

errors) of an image is a more stable measure than the root mean squared error to capture the total predictive capacity of the network ensemble. Its average over the entire test set equals 20 nm.

The relationship between the topography uncertainty and the root mean squared error of the topography predictions by the ensemble is shown for the test data set in figure 6. The uncertainty grows as the root mean squared error increases. A more detailed analysis of the uncertainty estimation is given in figure 7, where the profile of a topography is plotted along with the profile of the ensemble prediction and the estimated *uncertainty tube*, i.e. the profiles of ΔT, $f_\phi(\Delta L)$ and $f_\phi(\Delta L) \pm 1.96uc(f_\phi(\Delta L))$. The ground truth (in red) rarely leaves the uncertainty tube (in blue) and at the same time, the uncertainty tube is not too wide. The uncertainty tube is, in general, widest at the borders of the topography and smallest at its center. This behavior is in accordance with the corresponding sizes of rms errors. Some further examples are given in appendix C, figure 17.

The factor 1.96 for the uncertainty tube equals the 97.5% quantile of the standard normal distribution. This choice implies that, if the errors are normally distributed around the prediction of the ensemble with a standard deviation equal to $uc(f_\phi(\Delta L))$, then the uncertainty tube will encompass the difference between the predictions and the ground truth in 95% of the cases. We found the assumption of normally distributed errors to be justified in this application.

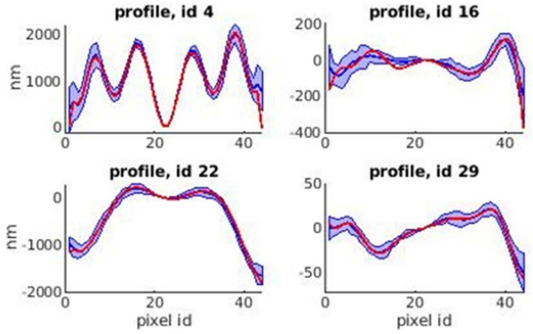

Figure 7. The profiles of four example topographies (cf figure 16) are plotted in red, along with the profiles of the ensemble predictions and the estimated uncertainty tubes in blue.

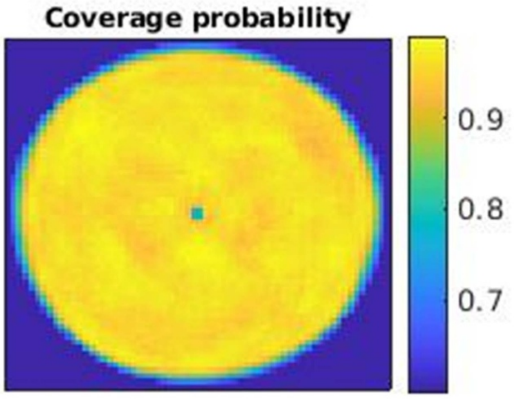

Figure 8. The uncertainty estimation covers the prediction error well if the coverage probability is close to one.

Figure 8 shows the coverage probability of the uncertainty estimation. For each pixel p of the topography, the *coverage probability* cp is estimated as follows:

$$cp(p) = \frac{1}{N}\sum_{i=1}^{N} g(u_i(p), y_i(p)),\; g(u_i, y_i) := \begin{cases} 1, & if \frac{|y_i - \hat{y}_i|}{1.96 u_i} \leq 1 \\ 0, & \text{otherwise} \end{cases}, \tag{5}$$

where y_i is the ground truth topography height at pixel p of the ith data sample, $\hat{y}_i$ is the corresponding predicted topography height and u_i the uncertainty estimate as defined in (4). The coverage probability indicates how likely it is that the ground truth is found around the prediction within the chosen uncertainty tube. Figure 8 shows the pixelwise coverage probabilities, which indicate that the calculated uncertainties characterize the size of the errors of the predictions well. The *total coverage probability* can be defined as follows:

$$\frac{1}{D'}\sum_{d=1}^{D'} \frac{1}{N}\sum_{i=1}^{N} g(u_i(p_d), y_i(p_d)) \in \mathbb{R}. \tag{6}$$

The observed total coverage probability equals 94%, which fits the intended 95% coverage probability well.

Altogether, the network ensemble makes good predictions and provides a trustworthy uncertainty estimate, not only per image, but also pixelwise, for the perfectly calibrated optical system.

5.2. Systematically introduced calibration error

In the last subsection, training and test data were generated by means of the same optical system. However, the network ensemble should also make trustworthy predictions on out-of-distribution data. Indeed, in any real measurement scenario, there will remain a calibration error. Therefore, the quality of the ensemble prediction and its uncertainty quantification are analyzed under the influence of a systematically introduced, growing calibration error. To this end, the optical system generating the test data is increasingly deviated from the optical system used to produce the training data. The chosen topographies in the test set remain the same (section 3).

Table 1. The influence of the introduced calibration error is analyzed on the test data set. The rows contain the following values: the percentage of induced calibration error, its impact on the input data expressed as the root mean squared error, the averaged root mean squared error over the single network predictions, the root mean squared error of the ensemble prediction, the median absolute error of the ensemble prediction, the topography uncertainty and the total coverage probability.

Calib error (in %)	0	10	20	30	40	50	60	70	80	90	100
Flawed input (in nm)	0	28	55	75	96	116	137	156	176	198	219
rmse single net (in nm)	95	137	178	229	280	331	383	435	484	538	596
rmse ensemble (in nm)	77	104	145	194	241	288	336	384	429	479	536
Median error (in nm)	20	29	42	55	68	82	95	110	125	139	154
Topography uc (in nm)	55	90	108	131	155	178	201	225	246	267	285
Total *cp* (in %)	94	94	84	78	75	73	71	70	69	68	67

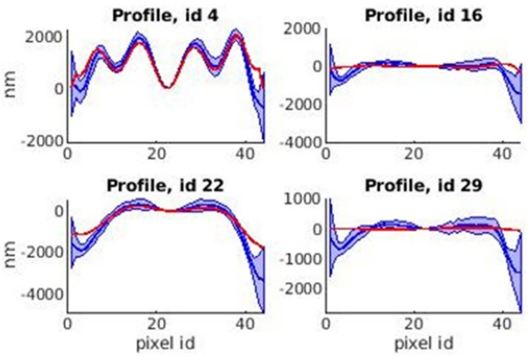

Figure 9. The profiles of four example topographies (figure 16) are plotted (red), along with the ensemble predictions and the estimated uncertainty tubes (blue) under full calibration error.

A first overview of the results is summarized in table 1. The calibration error is induced stepwise from zero to one hundred per cent, cf first row. The second row shows the actual impact of the corrupted optical system on the resulting optical path length differences, which are used as input data by the trained networks in order to predict the sought topography. Then, the root mean squared error of the ensemble prediction is given in the fourth row, which achieves consistently better results than the prediction of a single trained network (third row). A more robust measure of the ensemble prediction is the median error in the fourth row, because it is more stable against outliers at the edges of the topographies. The fifth row displays the mean topography uncertainties. Finally, the total coverage probabilities are calculated in the last row (cf equation (6)).

Analogous to figure 7, the profiles of the four test topographies are plotted (in red) in figure 9, along with their ensemble predictions and their estimated uncertainty tubes (in blue) for the full calibration error. In the first column, the ground truth topographies mostly are within the range defined by the predicted topographies and their calculated uncertainties. Furthermore, the ensemble predictions recognize the basic shapes of the sought topographies, except for their edges. In contrast, the topographies from the second column are not recognized well. Instead, the ensemble predictions resemble one another and predict a topography with much stronger peak to valley variability. This prediction behavior can be explained when considering the ground truth. The impact of the calibration error dominates the ensemble prediction for small topographies, while it has a smaller impact on larger difference topographies. More examples and the stepwise change in the ensemble prediction and its uncertainty estimation for the increasing calibration error are shown in the appendix C in figures 18–20, respectively.

The dependency between the ensemble uncertainty and the increasing calibration error is plotted in figure 10. The ensemble uncertainty grows with the growing calibration error, which is a desirable behavior for trustworthy predictions. Figure 10 shows the results over the entire test data set, while figure 9 presents the pixelwise results. In contrast, figure 11 shows the obtained results on the image level, where the root mean squared error of the ensemble prediction is plotted against the topography uncertainty for the different degrees of induced calibration error (also cf. figure 21). Again, the estimated uncertainty correlates well with the prediction error (as well as with the calibration error).

Finally, the coverage and total coverage probabilities (equations (5) and (6)) are shown in figure 12 in relation to the growing calibration error. Although the coverage probability slowly decreases, the total coverage probability stays at 94% after having induced 10% of the calibration error and still correctly covers

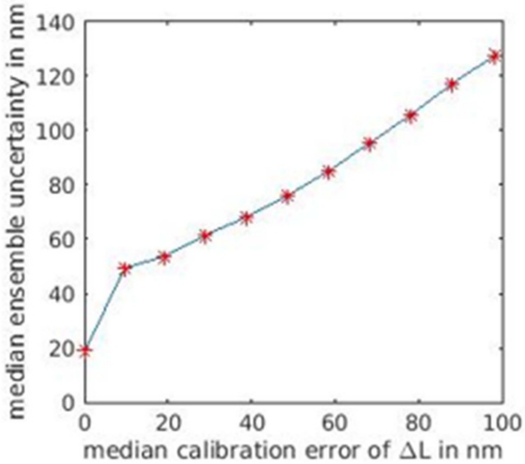

Figure 10. The median of the ensemble uncertainty on the test data set is plotted against the median of the absolute calibration error on the input data ΔL.

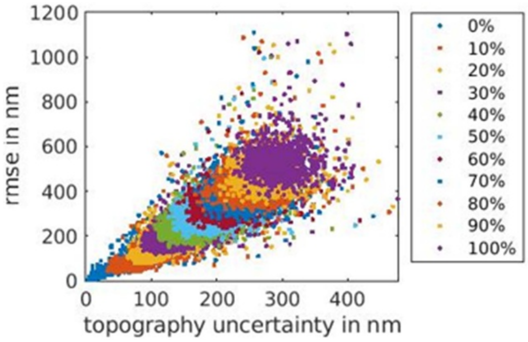

Figure 11. The root mean squared error of the ensemble prediction is plotted against the topography uncertainty. Each color represents a different amount of the introduced calibration error.

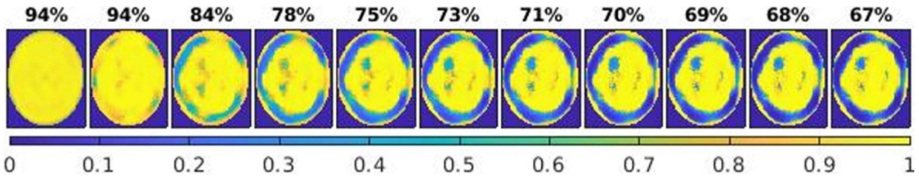

Figure 12. The coverage probabilities per pixel are plotted together with their total coverage probability on the test set for increasing calibration error (from left to right).

two thirds of the pixels for the maximal induced corruption of the input data in the test set. Furthermore, not the center but the topography edges are less well covered with increasing calibration error.

In sum, ensemble prediction and its uncertainty quantification are best when the optical system is perfectly calibrated and get worse as the quality of the calibration decreases. However, the uncertainty increases with a growing calibration error and appears to still reliably characterize the size of the errors in the predictions.

5.3. Noisy data

The previous subsection examines how systematically deviating the computer model of the optical system, that is used to simulate the test data, influences the network ensemble. Another source of error is noise in the data. Therefore, the ensemble prediction and its uncertainty estimation are analyzed in the following using noisy input data. Again, only the test data are modified, while the trained ensemble stays fixed.

Random white noise is added pixelwise to the input data with a standard deviation of 10 and 50 nm, respectively. This is done for the test data generated by the perfectly calibrated optical system as well as for

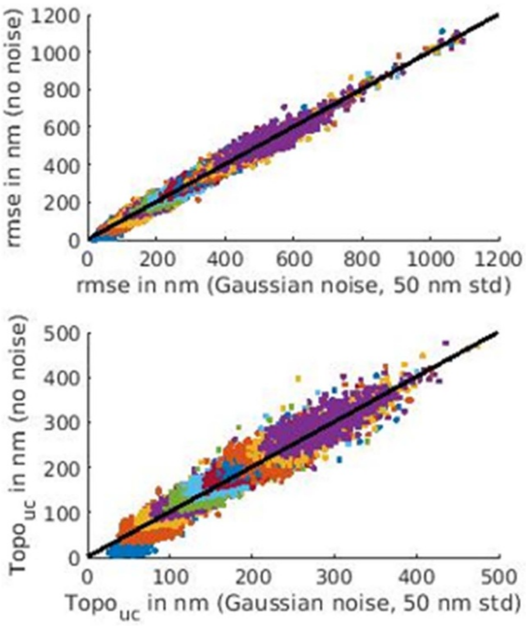

Figure 13. The root mean squared errors and topography uncertainties are plotted when feeding the network ensemble test data with and without additional Gaussian noise. The colors represent the amount of introduced calibration error in analogy to figure 12.

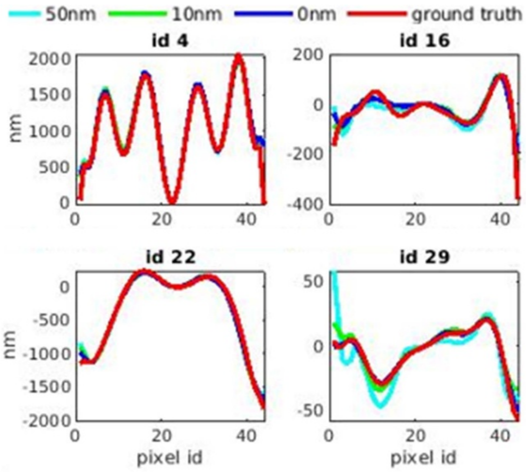

Figure 14. The ensemble predictions and their ground truth are plotted for four different test topographies under the influence of varying Gaussian noise (legend shows standard deviations) for the perfectly calibrated system.

the test data produced by the optical system after corrupting them with an increasing calibration error. Figure 13 shows the previous results plotted against the results when additionally corrupting the input data with normally distributed noise with a standard deviation of 50 nm. The root mean squared error of the ensemble predictions is plotted together with the topography uncertainties for the stepwise increasing introduced calibration error. There is no sudden loss in the prediction capability or the uncertainty estimation to be seen for noisy input data. Only for the perfectly calibrated case (dark blue) are the errors slightly larger for the disrupted data. However, also the topography uncertainties are slightly larger in this case which is a desirable behavior.

A closer look at some example topographies is given in figures 14 and 15, where the profiles are plotted. The former figure shows the ground truth together with the ensemble prediction for the perfectly calibrated data with and without noise. The noisy data have almost no impact on the ensemble prediction for the large topographies in the first row. In contrast, especially the noisy data with a standard deviation of 50 nm have a

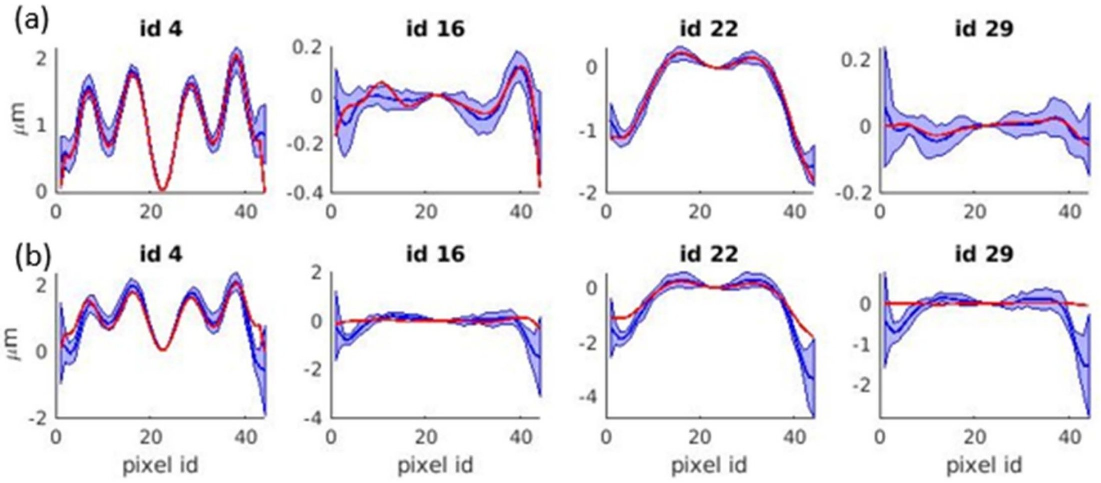

Figure 15. The profiles of the ensemble prediction and uncertainty tube are plotted (blue) with the underlying ground truth (red) for perturbed input data by white noise with a standard deviation of 50 nm for (a) the perfectly calibrated system and (b) the full calibration error.

visible influence on the ensemble prediction of the smaller topography at the bottom right. Nonetheless, the main shape of the topography is still recognized.

Figure 15 shows the ground truth along with the ensemble prediction and estimated uncertainty for the noisy input data with a standard deviation of 50 nm. The first row (a) shows the results for the perfectly calibrated optical system and can be compared to figure 7. Here, the uncertainties are much larger for the small topography (id 29) under the influence of noise, which is an appropriate behavior, as the prediction is also most strongly influenced by the noise for small topographies (figure 14). The second row (b) shows the same plots but for the maximal calibration error, which seems to dominate the noise; there is no visible difference between the plots in (b) and those in figure 9.

In total, the network ensemble is not largely affected by noisy data, even though it did not encounter them during training. The prediction error might increase in some cases, but then the uncertainty estimate rises as well.

6. Discussion and conclusion

In this paper, we have shown that ensemble learning is a powerful tool for reliably solving the inverse problem of reconstructing topographies from given optical path length differences. Moreover, the proposed ensemble method provides a sensible uncertainty quantification to its results, which is shown on the pixel level as well as on the image level. This is true not only for a diverse test set that is disjoint from the training data, but also under the influence of different additional error sources. The network ensemble was successfully tested on out-of-distribution data realized by systematically deviating the computer modeled optical system used to simulate the test data, which corresponds to an increasing calibration error. The estimated uncertainty grows in relation to the introduced calibration error and the prediction error, respectively. Furthermore, the ensemble behavior is analyzed under the influence of noisy input data using different amounts of white noise. The noisy input has almost no effect on the prediction. An impact can be seen mainly for small topographies. However, the uncertainty estimation is able to reflect this behavior as well.

From an application point of view, ensemble learning can be used to reliably solve the considered inverse topography reconstruction problem up to a certain accuracy faster than applied conventional methods after having trained the neural networks once for a specific design topography. Additionally, the network ensemble provides a consistent uncertainty quantification. Including other error sources such as positioning errors of the specimen or applying the proposed method to real data could be the next steps. Ensemble learning could also be applied to validate the current status of a calibration using a fixed reference specimen. Even if the topography is not perfectly known, the ensemble uncertainty should increase if the calibration worsens over time. In this way, the presented uncertainty estimate could be used to detect shifts in the calibration.

From a machine learning perspective, this application exemplifies the ability of ensemble methods to make trustworthy predictions and to provide an uncertainty quantification. The great potential of ensemble methods lies in their simple and straightforward implementation when solving high-dimensional problems.

The proposed uncertainty quantification mainly considers epistemic uncertainty; future work could address the explicit incorporation of aleatoric uncertainty sources as well. Also, establishing a high-dimensional benchmark data set to test and compare scalable uncertainty methods is referred to as future work.

Acknowledgment

The authors thank Manuel Stavridis for providing the simulation toolbox SimOptDevice and Michael Schulz for fruitful discussions about optical form measurements.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

Appendix A. Reference planes

The light path through the optical system is deviated by introducing the virtual reference planes R_1 and R_2 before and after the topography, respectively (figure 2). Note that two reference planes are required in order to ensure that the calibration is valid regardless of the surface under test.

Each reference plane is parameterized by a double fit of Zernike polynomials [4]. The parameterization of the *source* reference plane R_1 depends on the intersection (u, v) with a beam and its originating light source (U, V), i.e.:

$$L_{R_1}(u,v,U,V) = \sum_{i=1}^{I}\left(\sum_{j=1}^{J} Q_{ij} Z_j(U,V)\right) Z_i(u,v), \tag{7}$$

where $(Q_{ij})_{ij}$ is the matrix of Zernike coefficients with dimension $I \times J$. Analogously, the *pixel* reference plane R_2 depends on the intersection (m, n) with a beam and its corresponding pixel on the CCD (M, N), i.e.:

$$L_{R_2}(m,n,M,N) = \sum_{k=1}^{K}\left(\sum_{h=1}^{H} P_{kh} Z_h(M,N)\right) Z_k(m,n), \tag{8}$$

where $(P_{kh})_{kh}$ is the matrix of Zernike coefficients with dimension $K \times H$. The optical path lengths corresponding to the two reference planes are independent of the topography T, while the optical path length between the two reference planes depends on the given topography. The optical path lengths of the source to the source reference plane L_{R_1} and the pixel reference plane to the CCD L_{R_2} are added to the optical path length between the two reference planes L_T. Finally, the total optical path length difference is computed by subtracting the optical path length from the reference arm L_R (figure 1) from the resulting sum:

$$\begin{aligned} L(u,v,m,n,U,V,M,N,T) = {} & L_{R_1}(u,v,U,V) + L_{R_2}(m,n,M,N) \\ & + L_T(u,v,m,n,T) - L_R(M,N). \end{aligned} \tag{9}$$

The Zernike coefficients of the parameterized reference planes are iteratively adjusted during the calibration procedure [4]. This is done in such a way that the measurements performed using the computer modeled optical system increasingly resemble the measurements obtained from the real optical system for some well-known spherical calibration specimens. The calibration error is chosen to demonstrate effects of out-of-distribution data and does not necessarily reflect real world calibration errors [43].

Appendix B. Asphere

The asphere used as the design topography for the data generation process is quantified as follows. The aspherical coordinates $(A_4, A_6, \ldots, A_{16})^T$ are $(5.4145 \times 10^{03}\ \mathrm{m}^{-3}, -8.0413 \times 10^{05}\ \mathrm{m}^{-5}, -2.9871 \times 10^{09}\ \mathrm{m}^{-7}, -1.4918 \times 10^{12}\ \mathrm{m}^{-9}, 1.3777 \times 10^{15}\ \mathrm{m}^{-11}, 4.4258 \times 10^{18}\ \mathrm{m}^{-13}, -3.4928 \times 10^{21}\ \mathrm{m}^{-15})^T$, the conic constant κ equals -1 and the paraxial surface radius R equals 0.0202 m. The aspherical equation is given in [9] (2.2.2.1).

Appendix C. Additional plots and results

Table 2. This table shows how the results change with the number of networks in the ensemble in terms of the prediction capability (second row) and how good the estimated uncertainty covers the prediction errors (third row). The results are produced on the perfectly calibrated test data set without any additional noise.

Number of ensemble members	1	2	3	4	5	6	7	8
Median absolute error in nm	35	29	25	24	23	22	21	20
Coverage probability in %	0	60	77	84	88	91	93	94

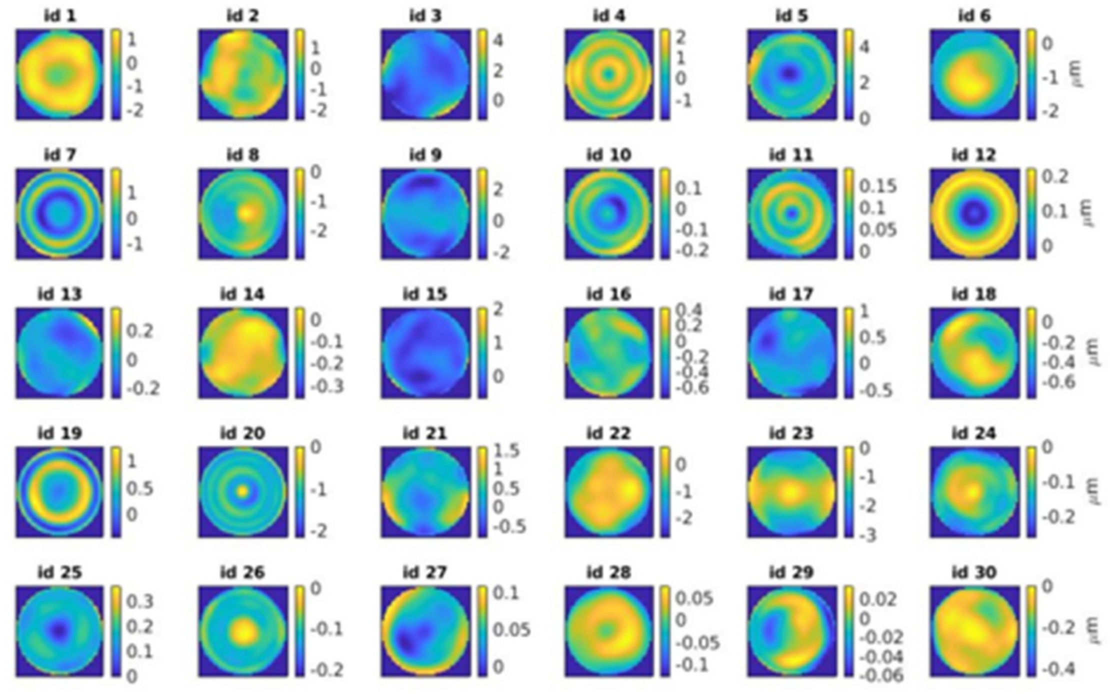

Figure 16. Some examples of generated test topography deviations ΔT are shown. The aperture has a radius of about 15 mm.

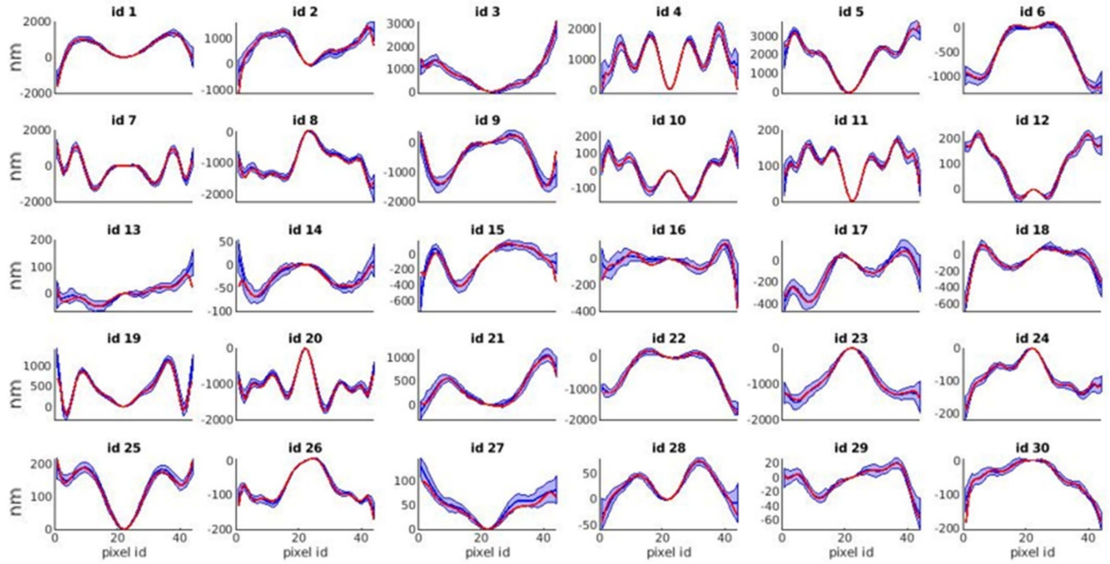

Figure 17. The profile plots of all topographies from figure 16 are plotted in red, while the ensemble predictions and the estimated uncertainties are plotted in blue.

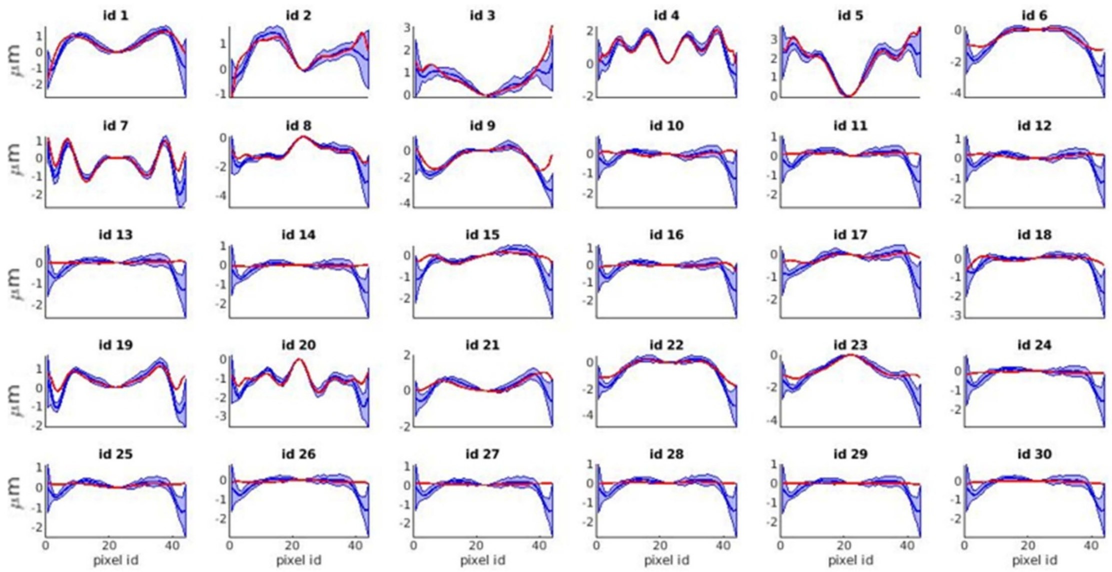

Figure 18. The profile plots of all topographies from figure 16 are plotted in red, while the ensemble predictions and the estimated uncertainties are plotted in blue for the full calibration error.

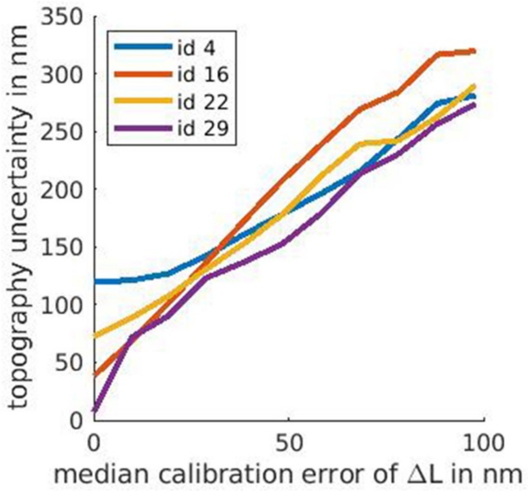

Figure 19. The topography uncertainty is plotted against the increasing calibration error for four individual difference topographies.

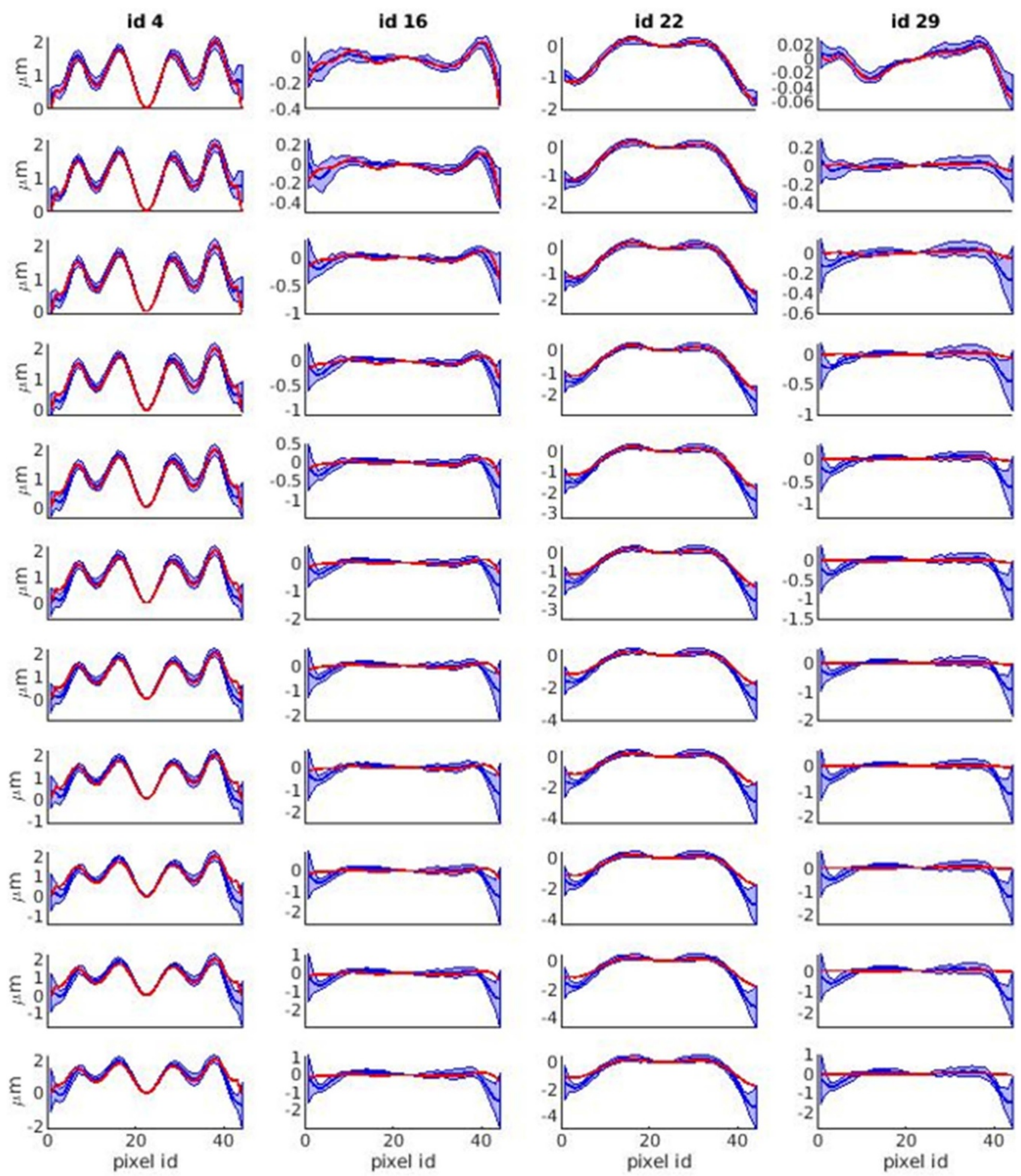

Figure 20. The profiles of the ensemble prediction and its estimated uncertainty tube (blue) are plotted together with the ground truth (red) for four difference topographies. The first row shows the results for the perfectly calibrated system. The calibration error is then increased by 10% in each row.

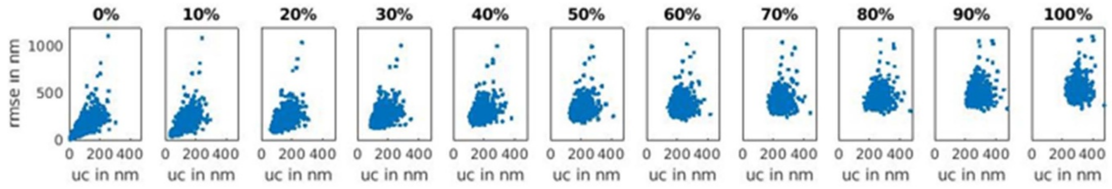

Figure 21. The root mean squared errors are plotted against the topography uncertainties of all data points in the test set under the influence of the systematically growing calibration error from left to right.

ORCID iD

Lara Hoffmann https://orcid.org/0000-0002-6211-413X

References

[1] Adadi A and Berrada M 2018 Peeking inside the black-box: a survey on explainable artificial intelligence (XAI) *IEEE Access* **6** 52138–60

[2] Akhtar N and Mian A 2018 Threat of adversarial attacks on deep learning in computer vision: a survey *IEEE Access* **6** 14410–30

[3] Ashukha A, Lyzhov A, Molchanov D and Vetrov D 2020 Pitfalls of in-domain uncertainty estimation and ensembling in deep learning (arXiv:2002.06470)

[4] Baer G, Schindler J, Pruss C, Siepmann J and Osten W 2014 Calibration of a non-null test interferometer for the measurement of aspheres and free-form surfaces *Opt. Express* **22** 31200–11

[5] Baer G, Schindler J, Pruss C, Siepmann J and Osten W 2014 Fast and flexible non-null testing of aspheres and free-form surfaces with the tilted-wave-interferometer *Int. J. Optomechatronics* **8** 242–50

[6] Baer G B 2017 *Ein Beitrag zur Kalibrierung von Nicht-Null-Interferometern zur Vermessung von Asphären und Freiformflächen* vol 86 (Stuttgart: Institut für Technische Optik, Universität Stuttgart)

[7] Baldi P and Sadowski P J 2013 Understanding dropout *Advances in Neural Information Processing Systems* vol 26, ed C J C Burges, L Bottou, M Welling, Z Ghahramani and K Q Weinberger (Curran Associates, Inc.) pp 2814–22

[8] Barbastathis G, Ozcan A and Situ G 2019 On the use of deep learning for computational imaging *Optica* **6** 921–43

[9] Braunecker B, Hentschel R and Tiziani H J 2008 *Advanced Optics Using Aspherical Elements* vol 173 (Bellingham: SPIE Press)

[10] Caldeira J and Nord B 2020 Deeply uncertain: comparing methods of uncertainty quantification in deep learning algorithms (arXiv:2004.10710)

[11] Chang T, Ryu D, Jo Y, Choi G, Min H-S and Park Y 2020 Calibration-free quantitative phase imaging using data-driven aberration modeling *Opt. Express* **28** 34835–47

[12] Dietterich T G 2000 Ensemble methods in machine learning *Int. Workshop on Multiple Classifier Systems* (Springer) pp 1–15

[13] Esser P and Sutter E 2018 A variational u-net for conditional appearance and shape generation *Proc. Conf. on Computer Vision and Pattern Recognition (Los Alamitos, CA, USA)* (IEEE Computer Society) pp 8857–66

[14] Fortmeier I *et al* 2020 Round robin comparison study on the form measurement of optical freeform surfaces *J. Eur. Opt. Soc.* **16** 1–15

[15] Fortmeier I, Stavridis M, Elster C and Schulz M 2017 Steps towards traceability for an asphere interferometer *Optical Measurement Systems for Industrial Inspection X* vol 10329 ed P Lehmann, W Osten and A A Gonçalves Jr (International Society for Optics and Photonics, SPIE) pp 790–8

[16] Fortmeier I, Stavridis M, Wiegmann A, Schulz M, Baer G, Pruss C, Osten W and Elster C 2013 Sensitivity analysis of tilted-wave interferometer asphere measurements using virtual experiments *Modeling Aspects in Optical Metrology IV* vol 8789 (International Society for Optics and Photonics) p 878907

[17] Fortmeier I, Stavridis M, Wiegmann A, Schulz M, Osten W and Elster C 2014 Analytical Jacobian and its application to tilted-wave interferometry *Opt. Express* **22** 21313–25

[18] Gal Y, Hron J and Kendall A 2017 Concrete dropout *Advances in Neural Information Processing Systems* vol 30 (Curran Associates, Inc.) pp 3581–90

[19] Goldstein R M, Zebker H A and Werner C L 1988 Satellite radar interferometry: two-dimensional phase unwrapping *Radio Sci.* **23** 713–20

[20] Gustafsson F K, Danelljan M and Schon T B 2020 Evaluating scalable Bayesian deep learning methods for robust computer vision *Proc. IEEE/ Conf. on Computer Vision and Pattern Recognition Workshops* pp 318–19

[21] Hansen L K and Salamon P 1990 Neural network ensembles *IEEE Trans. Pattern Anal. Mach. Intell.* **12** 993–1001

[22] Hoffmann L and Elster C 2020 Deep neural networks for computational optical form measurements *J. Sens. Sens. Syst.* **9** 301–7

[23] Hüllermeier E and Waegeman W 2021 Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods *Mach. Learn.* **110** 457–506

[24] Işıl Ç, Oktem F S and Koç A 2019 Deep iterative reconstruction for phase retrieval *Appl. Opt.* **58** 5422–31

[25] Kendall A and Gal Y 2017 What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems* vol 30 (Curran Associates, Inc.) pp 5574–84

[26] Kingma D P and Ba J 2014 Adam: a method for stochastic optimization (arXiv:1412.6980)

[27] Kingma D P, Salimans T and Welling M 2015 Variational dropout and the local reparameterization trick *Advances in Neural Information Processing Systems* vol 28 (Curran Associates, Inc.) pp 2575–83

[28] Kline S 1985 The purposes of uncertainty analysis *J. Fluids Eng.* **107** 153–60

[29] Kononenko I 1989 Bayesian neural networks *Biol. Cybern.* **61** 361–70

[30] Lakshminarayanan B, Pritzel A and Blundell C 2017 Simple and scalable predictive uncertainty estimation using deep ensembles *Advances in Neural Information Processing Systems* vol 30 (Curran Associates, Inc.) pp 6402–13

[31] Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W and Müller K-R 2019 Unmasking clever Hans predictors and assessing what machines really learn *Nat. Commun.* **10** 1096

[32] LeCun Y *et al* 1995 Convolutional networks for images, speech and time series *The Handbook of Brain Theory and Neural Networks* vol 3361 (Cambridge: MIT Press)

[33] Lee S, Purushwalkam S, Cogswell M, Crandall D and Batra D 2015 Why M heads are better than one: training a diverse ensemble of deep networks (arXiv:1511.06314)

[34] Li Z, Kovachki N, Azizzadenesheli K, Liu B, Bhattacharya K, Stuart A and Anandkumar A 2020 Fourier neural operator for parametric partial differential equations (arXiv:2010.08895)

[35] Martin J and Elster C 2020 Detecting unusual input to neural networks *Applied Intelligence* (Berlin: Springer) pp 1–12

[36] Nair V and Hinton G E 2010 Rectified linear units improve restricted Boltzmann machines *ICML*

[37] NTSB 2019 Highway investigation *National Transport Safety Board* HWY18MH010

[38] Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, Dillon J, Lakshminarayanan B and Snoek J 2019 Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift *Advances in Neural Information Processing Systems* vol 32 (Curran Associates, Inc.) pp 13991–4002

[39] Qin T 2020 *Deep Learning Basics* (Singapore: Springer) pp 25–46

[40] Ronneberger O, Fischer P and Brox T 2015 U-net: convolutional networks for biomedical image segmentation *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention* (Springer) pp 234–41

[41] Scalia G, Grambow C A, Pernici B, Li Y-P and Green W H 2020 Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction *J. Chem. Inf. Model.* **60** 2697–717

[42] Schachtschneider R, Stavridis M, Fortmeier I, Schulz M and Elster C 2019 SimOptDevice: a library for virtual optical experiments *J. Sens. Sens. Syst.* **8** 105–10

[43] Schindler J 2020 *Methoden zur Selbstkalibrierenden Vermessung von Asphären und FreiformfläChen in der Tilted-Wave-Interferometrie* vol 105 (Stuttgart: Institut für Technische Optik, Universität Stuttgart)

[44] Schindler J, Pruss C and Osten W 2019 Simultaneous removal of nonrotationally symmetric errors in tilted wave interferometry *Opt. Eng., Bellingham* **58** 074105

[45] Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D and Batra D 2017 Grad-CAM: visual explanations from deep networks via gradient-based localization *Proc. IEEE Int. Conf. Computer Vision* pp 618–26

[46] Sun Y, Wu M, Ruan W, Huang X, Kwiatkowska M and Kroening D 2018 Concolic testing for deep neural networks *Proc. 33rd ACM/IEEE Int. Conf. on Automated Software Engineering* pp 109–19

[47] Tian Y, Pei K, Jana S and Ray B 2018 DeepTest: automated testing of deep-neural-network-driven autonomous cars *Proc. 40th Int. Conf. on Software Engineering, ICSE '18 (New York, NY, USA)* (Association for Computing Machinery) pp 303–14

[48] Wang J and Silva D E 1980 Wave-front interpretation with Zernike polynomials *Appl. Opt.* **19** 1510–8

[49] Wang S, Cao J and Yu P 2020 Deep learning for spatio-temporal data mining: a survey *IEEE Trans. Knowl. Data Eng.* 1

[50] Xu K, Li J, Zhang M, Du S S, Kawarabayashi K-I and Jegelka S 2020 How neural networks extrapolate: from feedforward to graph neural networks (arXiv:2009.11848)

[51] Yao J, Pan W, Ghosh S and Doshi-Velez F 2019 Quality of uncertainty quantification for Bayesian neural network inference (arXiv:1906.09686)

[52] Young T, Hazarika D, Poria S and Cambria E 2018 Recent trends in deep learning based natural language processing *IEEE Comput. Intell. Mag.* **13** 55–75

[53] Zhang L, Zhou S, Li J and Yu B 2019 Deep neural network based calibration for freeform surface misalignments in general interferometer *Opt. Express* **27** 33709–23