



Extended range and aberration-free autofocusing via remote focusing and sequence-dependent learning

JIAHE CUI,^{1,4}  RAPHAËL TURCOTTE,²  NIGEL J. EMTAGE,³ AND MARTIN J. BOOTH^{1,5} 

¹Department of Engineering Science, University of Oxford, Parks Road, Oxford, OX1 3PJ, UK

²Tech4Health Institute, NYU Langone Health, New York, NY 10010, USA

³Department of Pharmacology, University of Oxford, Mansfield Road, Oxford, OX1 3QT, UK

⁴jiahe.cui@eng.ox.ac.uk

⁵martin.booth@eng.ox.ac.uk

Abstract: Rapid autofocusing over long distances is critical for tracking 3D topological variations and sample motion in real time. Taking advantage of a deformable mirror and Shack-Hartmann wavefront sensor, remote focusing can permit fast axial scanning with simultaneous correction of system-induced aberrations. Here, we report an autofocusing technique that combines remote focusing with sequence-dependent learning via a bidirectional long short term memory network. A 120 μm autofocusing range was achieved in a compact reflectance confocal microscope both in air and in refractive-index-mismatched media, with similar performance under arbitrary-thickness liquid layers up to 1 mm. The technique was validated on sample types not used for network training, as well as for tracking of continuous axial motion. These results demonstrate that the proposed technique is suitable for real-time aberration-free autofocusing over a large axial range, and provides unique advantages for biomedical, holographic and other related applications.

© 2021 Optical Society of America under the terms of the [OSA Open Access Publishing Agreement](#)

1. Introduction

Autofocusing to automatically acquire in-focus images is of wide interest for applications such as whole slide imaging [1,2], digital holography [3], and surgical microscopy [4]. Whole slide imaging systems typically use a dry high numerical aperture (NA) objective for imaging thin tissue slices. The main challenge lies in the need to acquire high quality images while dynamically tracking 3D topological variations of the sample surface over large lateral regions [1]. Conventionally, a set of axial-stack (z -stack) images is taken at different positions and a figure of merit is calculated for each image to find the focal plane corresponding to the peak or valley of the figure of merit [2]. This operation is time-consuming as the number of images required in each z -stack rapidly increases for a higher positioning accuracy [2,5]. In addition, the range is constrained to less than 40 μm due to the lack of structural details at severely out-of-focus positions where meaningful figure of merit calculations cannot be obtained [5,6]. To increase the autofocusing range, approaches that use dual-pinhole masks before the sensor [7], dual-LED illumination [8], as well as image-based wavefront sensing [9,10] have been introduced, successfully accommodating for defocus ranges above 80 μm . Reflection-based methods that track the angle of reflectance from a laser can enable fast autofocusing, but the objective focus can only be maintained at a fixed axial position above the reference plane, such that topological variations of the sample surface cannot be examined, inherently limiting its scope of application [1].

More recently, machine learning methods applied to autofocusing have been particularly advantageous for enhancing speed and throughput [11–18], accuracy by means of combining

information from different domains [12], as well as robustness towards non-ideal scenarios such as sample tilt [15,17], cylindrical curvature [15], and unevenly distributed focal points [18]. For the majority of applications, convolutional neural network (CNNs) were used to interpret defocus information from acquired sample images or holograms [12–15,17,18]. An in-focus image was then acquired by either mechanical adjustment of the sample stage and objective lens [13], or virtual refocusing [15,17,18]. Despite their many advantages, deep learning methods possess some inherent drawbacks that cannot be neglected. First, networks that directly use sample images as input need to make interpretations from sufficient structural details, which dictates a relatively short defocus range of at most $\pm 20\ \mu\text{m}$ [11] and typically around $\pm 10\ \mu\text{m}$ [12,15,17,18]. To obtain a larger defocus range, structural information can be embedded within holograms using interference [14], or a Fourier transform can be performed on the sample image [16]. For imaging of thick tissue with severely tilted or unflat surfaces, as well as applications where sample movement is likely to occur [4], much larger defocus ranges are desired. Second, despite the short prediction time, a significant amount of computational power is required during the training stage on image data, which can typically take over a day even on powerful GPUs [11,12,14,17]. Third, networks that rely on information from specimen images can be difficult to generalise towards untrained sample types [16,18,19].

To address the above limitations while preserving merits of the network prediction speed, one possible solution is to extract positional information from sources other than sample images, and adopt less computationally intensive networks [16] that make use of non-image-based input data. Techniques proposed in [9,10] inferred phase information by solving the transport of intensity equation and demonstrated a much extended defocus range, though at the expense of using multiple correction loops. Alternatively, optical phase can be directly measured at the pupil plane using specialised devices such as a Shack-Hartmann wavefront sensor (SHWS) [20]. The latter eradicates dependency on structural details, which fundamentally overcomes limitations to the defocus range, while greatly enhancing generality towards different sample types. Apart from defocus, the SHWS can simultaneously measure other aberrations that degrade the image quality [21]. By applying an equal but opposite phase profile at the pupil plane using an adaptive element such as a deformable mirror (DM) or spatial light modulator, one can compensate for these aberrations through adaptive optics and restore the image quality [22]. The phase can be expanded into a set of Zernike polynomials [23] that are orthogonal over a unit circle [24], and the coefficients of these expansion terms, commonly referred to as Zernike coefficients, further reduce the representation of phase information from a 2D phase function into a short list of coefficients. Such data can be well interpreted by a much simpler network, such as a multilayer perceptron [16] or recurrent neural network. In the context of this work, we take advantage of a bidirectional long short term memory (Bi-LSTM) network [25,26], which is designed to learn from both directions of an input sequence for prediction of the output at the current data point, or shift, and extract maximum information from the sequential ordering [26]. It has high noise tolerance and requires less operations when updating weights that connect adjacent units for each shift, making it faster to train, easier to generalise, and less dependent on parameter fine tuning [25]. Finally, axial scanning by means of remote focusing (RF) can avoid slow movements of the sample stage and objective lens. Compared to other fast refocusing elements such as acousto-optic deflectors [27,28] and electrically tunable lenses [29], RF can compensate for system aberrations induced by the change of beam divergence at the back aperture of an infinity-corrected objective and enable aberration-free focus control [30–32].

In this work, we report an extended range autofocusing technique that combines RF with sequence-dependent learning via Bi-LSTM network prediction. A DM and SHWS were used to record phase information, as well as correct for system-induced aberrations during axial scanning. Three sequential shift measurements consisting of specific Zernike coefficients acquired at predetermined axial positions were used to predict RF voltages for autofocusing

(sequence-dependent learning). Experiments were performed both in air and refractive-index-mismatched media, such as arbitrary-thickness liquid layers up to 1 mm, in order to test validity for samples both exposed to air and immersed in a layer of physiological liquid. Network generality was evaluated for fields of view (FOVs) of different sizes, as well as sample types not seen during network training. Finally, proof-of-concept experiments were carried out for tracking of continuous axial motion in real time.

2. Methods

2.1. System setup

Imaging was performed on the basis of our previously developed compact reflectance confocal microscope described in [33]. For this work, a large stroke DM (DM-69, Alpao) with a settling time of 1 ms and comprising of 69 actuators was incorporated at a pupil plane conjugate to the back aperture of a 0.42 NA, 20 mm working distance dry objective lens (MY20X-804, Mitutoyo). A custom SHWS was positioned conjugate to the DM membrane in a separate wavefront sensing path. See Fig. 1 for the schematic diagram and Section 1 of Supplement 1 for detailed descriptions of the system modifications.

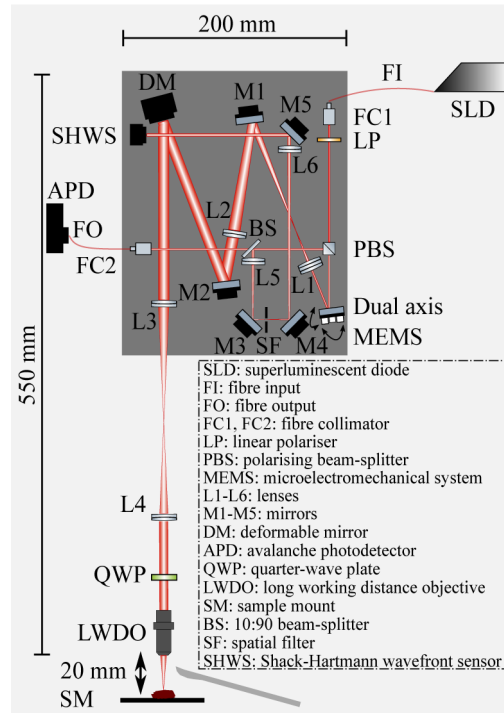


Fig. 1. Schematic diagram of autofocusing system incorporating a large stroke DM and a custom SHWS.

2.2. Remote focusing calibration

Voltages that control DM actuators to deform the membrane for fine axial refocussing were calibrated by repeatedly displacing a piece of white card (to avoid specular reflections) an amount Δd μm away from the objective natural focal plane using a sample translation stage, and compensating for this displacement using a closed-loop sensor-based adaptive optics scheme [34–36]. This allowed simultaneous correction of system aberrations induced at each calibration

step. In particular, a set of DM control voltages was obtained with the objective at its natural focal plane, which we call the DM system flat file. Linear interpolation was then performed to acquire voltages for finer axial steps. See Section 2 of Supplement 1 for details of the calibration procedure.

The Strehl ratio (SR), which can be defined as the ratio of intensity on-axis between an aberrated and unaberrated beam [37], is commonly used as a performance criterion for high resolution imaging systems. During RF calibration, the Strehl ratio after closed-loop adaptive optics correction at each calibration step was estimated using the first 69 detected Zernike coefficients via the following expression [24],

$$SR = \exp \left[- \left(\frac{2\pi}{\lambda} \right)^2 \cdot \left(\sum_{i=2}^N a_i^2 \right) \right], \quad (1)$$

where λ is the detection wavelength, a_i is the i th Zernike coefficient, and N is the total number of Zernike modes. 69 modes were included to account for high order modes that may be generated during closed-loop correction. The RF range was defined as the distance over which Strehl ratios maintained above 0.8 [37], resulting in a $\pm 60 \mu\text{m}$ range with negligible aberration.

2.3. Autofocusing workflow

The workflow of the proposed autofocusing scheme is illustrated in Fig. 2 for the typical case of surface tracking. Figure 2(a) depicts the main procedure for network training, including data collection and preprocessing, and pairing of input-ground truth data. Training data was collected by first applying a DM system flat file and adjusting the sample stage axially such that the tissue surface coincided with the objective natural focal plane. Then, RF voltages were applied to the DM for axial scanning with $0.5 \mu\text{m}$ steps from $-60 \mu\text{m}$ focal displacement to $+60 \mu\text{m}$. At each axial step z_d , phase aberration measurements were taken with the SHWS when scanning over a small lateral region ($< 20 \mu\text{m}$) to form one dataset. This was repeated multiple times on mouse skull exposed to air, as well as that beneath an arbitrary-thickness liquid layer within 1 mm. Various surface conditions were taken into account, such as regions with different surface roughness, curvature, and signal levels.

Collected datasets were then preprocessed by sampling at 3 sequential shifts τ_1 , τ_2 , τ_3 for each axial step. We consider sampling in the negative and positive directions as the past and future states, respectively. The full RF range was partitioned into 3 regions, namely the central, negative, and positive regions, as defined by stride length Δz . The central region incorporated axial steps within $[-60 + \Delta z, 60 - \Delta z] \mu\text{m}$ range of the tissue surface, the negative region $[-60, -60 + \Delta z] \mu\text{m}$, and the positive region $(60 - \Delta z, 60] \mu\text{m}$. For all 3 regions, τ_1 corresponded to the currently evaluated position z_d . In the case of the central region, τ_2 corresponded to an axial step Δz in the negative direction $z_d - \Delta z$, and τ_3 corresponded to an axial step Δz in the positive direction $z_d + \Delta z$. The Zernike coefficients as retrieved by the SHWS at each of the 3 axial positions $[a_{1,1}, a_{1,2}, \dots, a_{1,N}]$, $[a_{2,1}, a_{2,2}, \dots, a_{2,N}]$, and $[a_{3,1}, a_{3,2}, \dots, a_{3,N}]$, were then reorganised into one input training sample that consisted of 3 shifts and N features. For the negative region, τ_2 was sampled at $z_d + \Delta z$, while τ_3 corresponded to positions further away from z_d , denoted by $z_d + \Delta z + \Delta z_-$. For the positive region, τ_2 was sampled at $z_d - \Delta z$, while τ_3 was sampled at $z_d - \Delta z - \Delta z_+$. Finally, input training samples were paired with a set of ground truth RF voltages corresponding to the opposite amount of focal displacement $-z_d$, denoted as V_{-z_d} , to form one training sample. Separate networks of the same architecture were trained for each region.

Figure 2(b) demonstrates how the trained networks enable RF voltage prediction during real-time autofocusing. At any arbitrary axial position z_d , RF voltages corresponding to z_d , $z_d - \Delta z$, and $z_d + \Delta z$ were first sequentially applied to the DM. At each position, in parallel to image acquisition, Zernike coefficients retrieved by the SHWS were analysed to make a decision

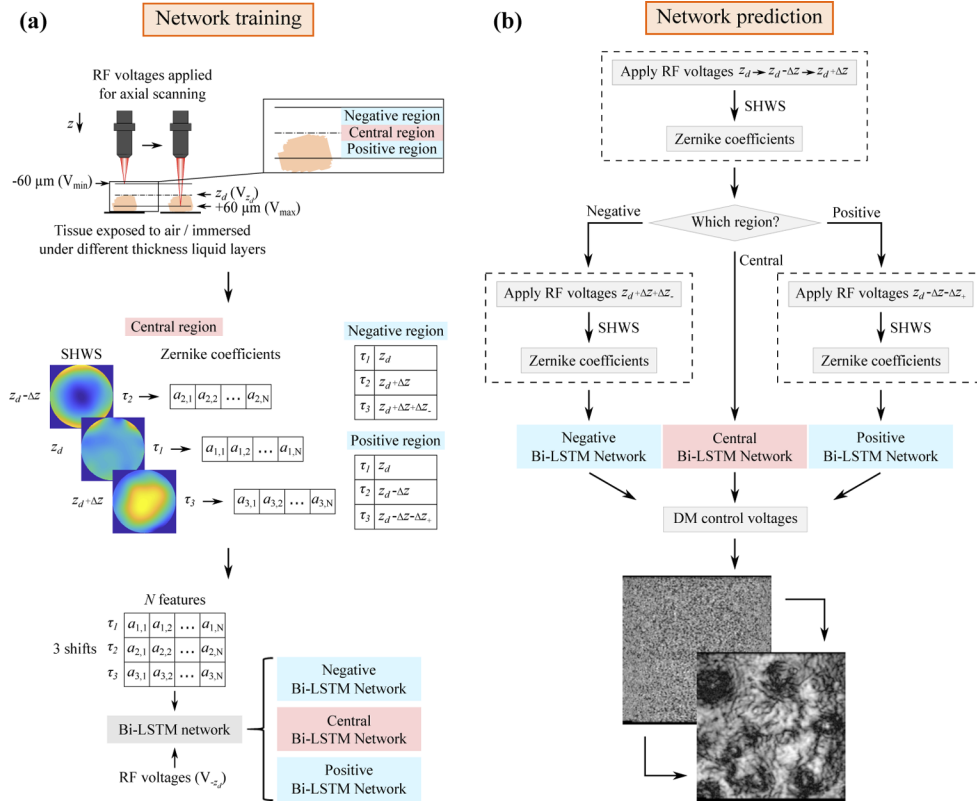


Fig. 2. Autofocusing workflow via RF and sequence-dependent learning. (a) During data collection, RF voltages V_{z_d} were first applied for axial scanning. At each axial step, Zernike coefficients $a_{j,i}$, where j represents the j th shift and i the i th Zernike coefficient, sequentially measured at 3 axial positions were used as network input and paired with RF voltages corresponding to the opposite amount of focal displacement. Separate Bi-LSTM networks of the same architecture were trained for the central, negative, and positive regions. (b) During experiments, Zernike coefficients measured at the same axial positions were passed into the corresponding Bi-LSTM network for direct prediction of DM control voltages.

based upon the consecutive measurements of Zernike defocus (Z_2^0) on which region the currently evaluated position fell within. In the case of the central, retrieved Zernike coefficients were then reorganised into a matrix of 3 shifts and N features as in the training stage and passed into the central Bi-LSTM network. For the negative region, a 4th phase measurement would be made at $z_d + \Delta z + \Delta z_-$, and these Zernike coefficients would be used along with those acquired at z_d and $z_d + \Delta z$ to form input data for the negative Bi-LSTM network. Similar procedures apply to the positive region at positions sampled during network training. Predicted RF voltages were then directly applied for refocusing. See Section 3 of [Supplement 1](#) for sample preparation procedures.

2.4. Evaluation of Zernike distribution curves

The distribution of primary Zernike modes over the full $\pm 60 \mu\text{m}$ defocus range as interpreted by the SHWS was evaluated for all datasets (Fig. 3). In addition to Zernike defocus (Z_2^0), which provides the most critical positional information, other low order Zernike modes such as primary spherical aberration (Z_4^0), vertical astigmatism (Z_2^2), horizontal coma (Z_3^1), vertical coma (Z_3^{-1}), and oblique astigmatism (Z_2^{-2}), can also be present in the system as aberrations introduced

during RF and were simultaneously assessed. A 90% data distribution interval was calculated to show how the majority of data was distributed, while the remaining 10% which possibly included outliers and extremely noisy data, were also used during network training to improve generality. A 95% confidence interval was obtained from the full dataset to represent the level of uncertainty in the curve estimation process, and the smoothed mean was also calculated to understand the overall distribution trend. More details of the distribution and confidence interval are given in Section 4 of Supplement 1. Results showed that despite the diversity of surface conditions and liquid layer thicknesses, Zernike coefficients could be inferred with a reasonably high accuracy throughout the full defocus range. It could be seen that although some of the curves were monotonic within a small defocus range, none of them were monotonic within the full defocus range. Therefore, in order to predict our precise location at an arbitrary point within the full defocus range, multiple shifts were required. During experiments, discrimination of the currently evaluated region was inferred from the 3 sequential measurements of Zernike defocus (Z_2^0) according to asymptotes of the distribution curve via the following criterion,

$$\begin{cases} \text{negative} & \text{if } a_3 \leq a_1 \leq a_2 \text{ and } a_2 < 0.1 \\ \text{positive} & \text{if } a_3 \leq a_1 \leq a_2 \text{ and } a_3 > 0 \\ \text{central} & \text{otherwise,} \end{cases} \quad (2)$$

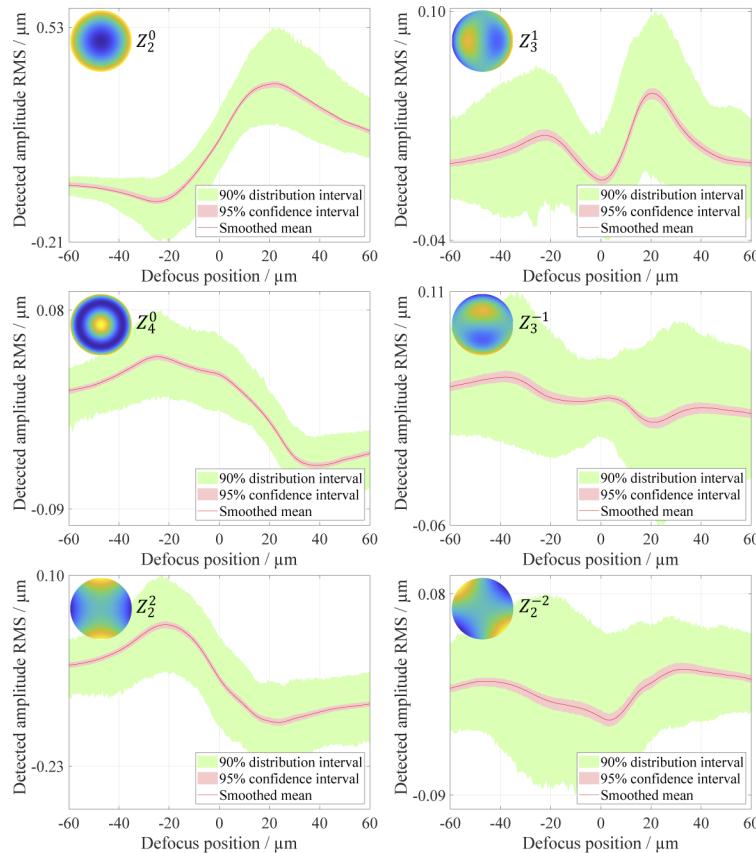


Fig. 3. Wavefront maps and distribution plots of primary Zernike modes as a function of defocus position incorporating data from all datasets.

where a_1 , a_2 , and a_3 represent measured Zernike defocus values at positions z_d , $z_d - \Delta z$, and $z_d + \Delta z$, respectively.

2.5. Bi-LSTM network

The Bi-LSTM network architecture is illustrated in Fig. 4. It comprises of 4 Bi-LSTM layers each with 64 units and a dense layer with 69 nodes for each of the 69 DM actuators. Input features were individually normalised between $[-1,1]$ and batch normalisation was performed before each Bi-LSTM layer. ReLU activation was used to introduce non-linearity in Bi-LSTM layers and linear activation was used in the dense layer for direct voltage prediction. Random weights were initialised with a He uniform kernel and the adaptive moment estimation (Adam) optimiser [38] was adopted to iteratively update weights and biases with a learning rate of 1×10^{-3} . The network was implemented using Keras with a Tensorflow backend.

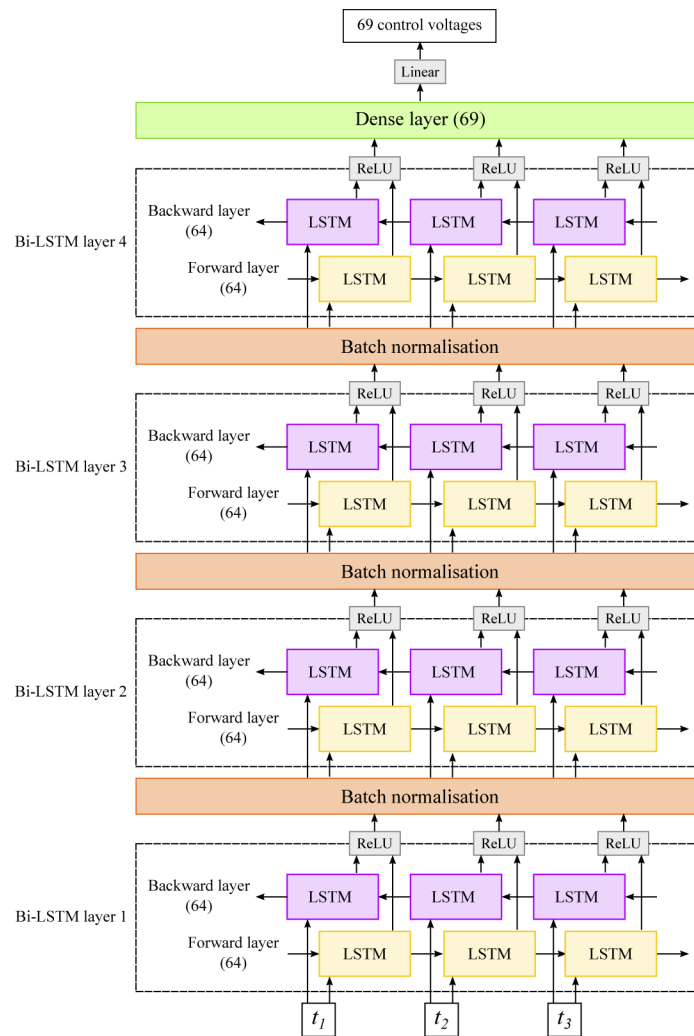


Fig. 4. Architecture of the proposed bidirectional long short term memory (Bi-LSTM) network comprising of 4 Bi-LSTM layers each with 64 units, batch normalisation, and a dense layer with 69 nodes to control 69 DM actuators. ReLU and linear activation functions were used in the Bi-LSTM layers and final dense layer, respectively.

Offline optimisation routines were performed to find the optimum stride lengths Δz , Δz_- , and Δz_+ , the number of Zernike modes to use as input features, as well as the batch size and number of datasets for network training (Fig. 5). Here we demonstrate validation results after all above parameters had been obtained such that only one parameter was interrogated at a time with other parameters set as the optimum value. 160 datasets were used in the first 3 tasks, 140 for network training with a validation split of 0.2, and 20 as test data for blind inference, which is process of predicting RF voltages with input data not used during the training stage. Samples were shuffled between epochs to prevent the model from learning order dependence between independent samples. In all cases, the root mean square error (RMS error) between predicted

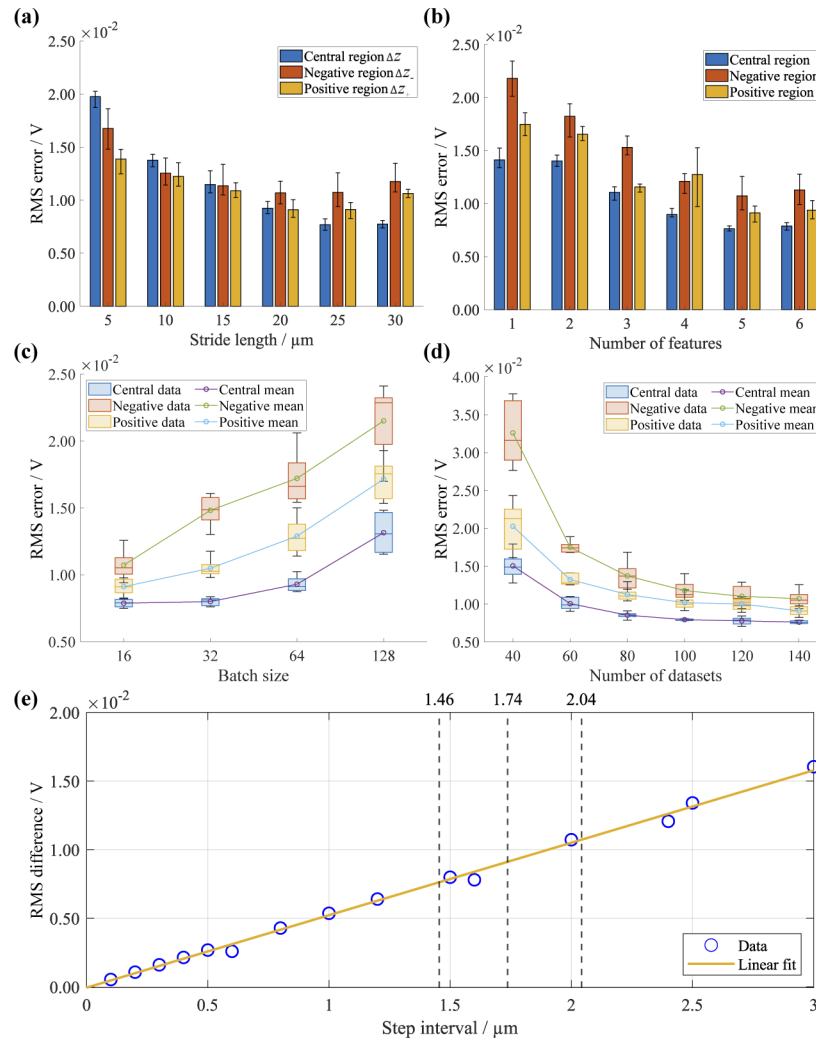


Fig. 5. Offline network optimisation results. (a) Test RMS error as a function of stride length Δz , Δz_- , and Δz_+ ; (b) Test RMS error versus the number of Zernike modes incorporated as input features; (c) Test RMS error as a function of batch size for each training epoch; (d) Test RMS error versus the number of datasets used for network training. (e) RMS difference between RF voltages corresponding to calibrated axial step intervals. Dashed vertical lines indicate the equivalent prediction error in terms of distance for the final RMS error during offline optimisation.

and ground truth RF voltages defined the loss function and was compared with the root mean square difference (RMS difference) between RF voltages corresponding to calibrated axial step intervals. The central Bi-LSTM network was empirically trained for 30 epochs and the positive and negative Bi-LSTM networks 70 epochs, while more epochs led to overfitting.

3. Results

3.1. Offline network optimisation results

To determine stride length Δz , the central Bi-LSTM network was individually trained using samples generated with stride length increments of $5\text{ }\mu\text{m}$ within the range of $5\text{ }\mu\text{m}$ to $30\text{ }\mu\text{m}$. The optimum stride length Δz_{op} corresponding to the smallest RMS error was then employed to find $\Delta z_{-\text{op}}$ and $\Delta z_{+\text{op}}$ using the negative and positive Bi-LSTM networks. As outlined in Section 2.3, Δz determined the size of the 3 defocus regions. Therefore, the amount of training and test samples varied with a fixed number of datasets during optimisation of Δz , while this number remained constant for that of Δz_{-} and Δz_{+} . The number of training and test samples used for optimisation of Δz , Δz_{-} , and Δz_{+} is given in Table S1. of Supplement 1, and the mean RMS error over 5 test runs for each examined stride length is plotted in Fig. 5(a) with corresponding error bars. Results suggested an optimum value of $25\text{ }\mu\text{m}$ for all stride lengths Δz_{op} , $\Delta z_{-\text{op}}$, and $\Delta z_{+\text{op}}$. A complete summary of the raw RMS error data is provided in Table S2. of Supplement 1. The optimum number of input features was determined by choosing combinations of the primary Zernike modes evaluated in Section 2.4 according to their azimuthal index orders and level of uncertainty as indicated by the 95% confidence interval. The combination of Zernike modes used in each case is given in Table 1. Figure 5(b) shows how the RMS error changed with increasing input features over 5 test runs for each of the 3 defocus regions. A summary of the raw RMS error data is provided in Table S3. of Supplement 1. Results showed that in all 3 cases, prediction accuracy was highest when 5 Zernike modes were included as input features. This conclusion suggested that the incorporation of other primary Zernike modes in addition to Zernike defocus can indeed provide useful information for enhancement of network performance. However, there is an extent to which more input features can be beneficial, dependent upon system-specific phase aberrations, and driving beyond this limit may lead to higher chances of overfitting to noisy data. The 3 network models were then trained with different batch sizes of 16, 32, 64, and 128. A monotonic decrease in RMS error was seen with smaller batch sizes in all 3 cases, suggesting an optimum value of 16 (Fig. 5(c)). Finally, networks were trained with increasing numbers of datasets to find the threshold for a sufficient prediction accuracy. Results in Fig. 5(d) showed that the RMS error dropped rapidly between 40~80 sets of data, while a moderate improvement was observed afterwards. Considering the trade-off between training time (effort required for data collection) and prediction accuracy, the mean RMS error after training with 140 datasets (0.0077 V, 0.0091 V, and 0.0107 V for the central, positive and negative defocus regions, respectively) was compared with the RMS difference between RF voltages corresponding to calibrated axial step intervals, as illustrated in Fig. 5(e). Equivalent prediction error in terms of distance was inferred from the linear fit to be $1.46\text{ }\mu\text{m}$, $1.74\text{ }\mu\text{m}$, and $2.04\text{ }\mu\text{m}$ for the central, positive, and negative regions, which was much smaller than half the system axial resolution of $5.5\text{ }\mu\text{m}$, indicating a sufficient accuracy according to the Nyquist criterion. Therefore, 140 datasets were used to generate final network models.

Networks were trained on a desktop computer with Intel 8-core i7-7820X 3.60 GHz CPU and 16 GB RAM, which took less than 5 mins for the central network, and 4 mins for the positive and negative networks. Real-time autofocusing was performed on a laptop with Intel 6-core i7-9750H 2.60 GHz CPU and 16 GB RAM. Blind inference took less than 0.037 s for all 3 networks. One full autofocusing cycle took 0.31 s when in the central region, and 0.40 s when in the positive and negative regions. Phase acquisition time could be flexibly adjusted.

Table 1. Combination of Zernike modes used for evaluation of the optimum number of input features.

Number of features		1	2	3	4	5	6
Zernike modes	Z_2^0	✓	✓	✓	✓	✓	✓
	Z_4^0		✓	✓	✓	✓	✓
	Z_2^2			✓	✓	✓	✓
	Z_3^1				✓	✓	✓
	Z_3^{-1}					✓	✓
	Z_2^{-2}						✓

3.2. Experimental results

The proposed autofocusing technique was first characterised on mouse skull exposed to air and beneath 250- μm , 500- μm , 750- μm , and 1000- μm thick liquid layers over the same FOV during data collection. With a DM system flat file applied, the sample was translated to defocused positions along the optical axis with 10 μm step intervals over the $\pm 60 \mu\text{m}$ range. At each position z_d , 5 independent autofocusing tests were performed, and the average image intensity I_d after refocusing to the sample surface was recorded for each instance. Considering the approximate relation of $h_{\text{eff}} = |h_a|^4$ between the effective system intensity point spread function h_{eff} and single path amplitude point spread function h_a in a reflectance confocal microscope [33], I_d was compared with the average image intensity I_{ref} acquired when no defocus was induced by calculating the relative Strehl ratio, which is proportional to $\sqrt{I_d/I_{\text{ref}}}$. Results showed that under all 5 scenarios, the trained Bi-LSTM network allowed tracking of the tissue surface with a relative Strehl ratio of above 0.8 for all examined positions (Fig. 6(a)), regardless of the liquid layer thickness. Narrow error bars were obtained, indicating a good level of repeatability. Using the relative Strehl ratio as a quality metric takes into account not only defocus information, but also image quality in the presence of optical aberrations.

Generalisation of the trained networks was evaluated over FOVs of different sizes and sample types not seen during network training. 50 μm and 100 μm FOVs were individually assessed on mouse skull exposed to air and fixed mouse brain. As illustrated in Fig. 6(b), autofocusing was performed at each 10 μm step from $-60 \mu\text{m}$ to $+60 \mu\text{m}$ of focal displacement by translation of the sample stage, and imaging was performed over a tilted region of the curved skull so as to simultaneously assess the network's resilience towards sample tilt. Exposure of the SHWS was set to 40 ms when scanning over 50 μm FOVs and 80 ms over 100 μm FOVs. Results at each 20 μm axial step are displayed in Fig. 6(c) for the case of autofocusing on mouse skull, and Fig. 6(d) for fixed mouse brain. Images were individually normalised between maximum and minimum pixel intensities for real-time display. Comparison was made between images before and after autofocusing, and the reference plane z_{ref} in each experiment when no defocus was introduced is highlighted in the magenta box. Results obtained after autofocusing from different defocus positions in both specimens showed comparable image quality to that obtained at the reference plane for both 50 μm and 100 μm FOVs. The focal plane remained stationary when autofocusing was performed without pre-induced defocus, and structural details at the tissue surface came into focus at all other positions, demonstrating robustness towards sample tilt, as well as validity for unseen specimens.

Next, through a series of 4 experiments we assessed the technique's ability to track continuous axial motion in real time. A 50 μm FOV was used in the first 3 experiments, and a 100 μm FOV in the final experiment. Imaging was performed at 4 fps and a predetermined number of consecutive images were recorded for each experiment. All Visualizations are played at 8 fps for demonstration purposes. The process of making sequential shift measurements was observed as brief 'blinks'. Videos of the same process without autofocusing were recorded in each case. The

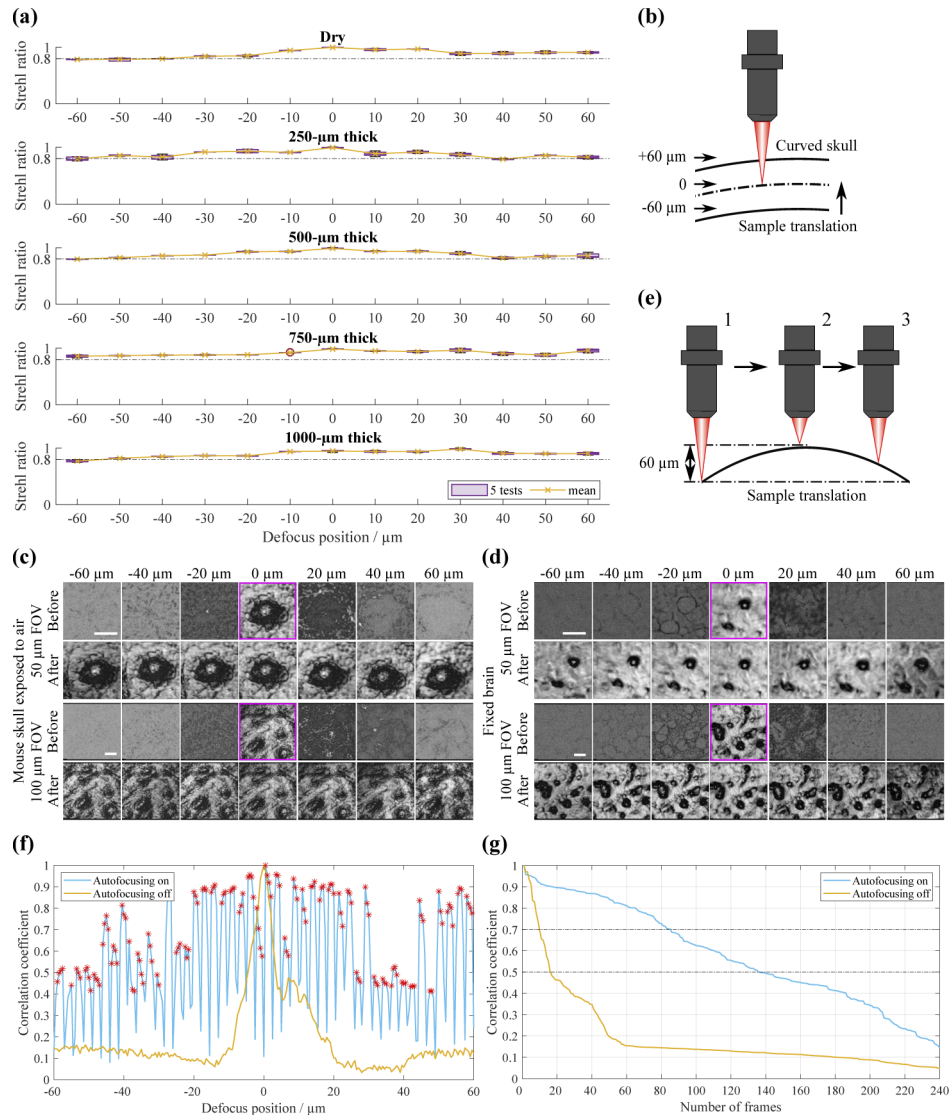


Fig. 6. Experimental validations using the proposed autofocus scheme. (a) Characterisation results on mouse skull exposed to air and that beneath 250- μm , 500- μm , 750- μm , and 1000- μm thick liquid layers over lateral scan regions smaller than 20 μm . 5 independent tests were performed at each 10 μm step interval over the full $\pm 60 \mu\text{m}$ range. (b) $-60 \mu\text{m}$ to $+60 \mu\text{m}$ focal displacement introduced by sample stage translation at tilted region of curved skull during validation of different size FOVs. (c)-(d) Validation results over different size FOVs and fixed specimens. 50 μm and 100 μm FOVs individually assessed with: (c) mouse skull exposed to air; (d) fixed mouse brain. Images before and after autofocus are displayed for comparison. Magenta box: reference image in each experiment. Scale bar: 20 μm . (e) Schematic of the relative motion between focal plane and sample surface during the course of Visualization 1. (f) Pearson correlation coefficients computed for each frame in Visualization 3 and Visualization 4 with respect to that acquired at the natural focal plane as a function of defocus position. Red asterisks: frames before (after) seeing a sudden decrease (increase) in correlation. (g) The number of frames in Visualization 3 and Visualization 4 with a Pearson correlation coefficient above each value. Dotted lines indicate a strong (>0.7) or moderate (>0.5) relationship.

first experiment showed how the proposed technique allowed continuous tracking of samples with severely tilted and unflat surfaces. With a DM system flat file applied, the natural focal plane was first positioned $60\text{ }\mu\text{m}$ deep below the peak of curved mouse skull. Next, the stage was translated laterally until a different surface region came into focus within the $50\text{ }\mu\text{m}$ FOV. This lateral position was recorded as the starting point for [Visualization 1](#) and [Visualization 2](#). Then in [Visualization 1](#), autofocusing was switched on for a feedback rate of 0.5 Hz , and the stage was translated back along the same path until it reached the skull peak, before surpassing it in the other direction, as illustrated in Fig. 6(e). From [Visualization 1](#) it could be appreciated that the proposed technique allowed tracking of the surface morphology after each autofocusing routine and continuous examination of structural details was enabled throughout the whole period. In comparison, without autofocusing the skull surface was lost after a few frames, as seen in [Visualization 2](#).

In the second and third experiment, autofocusing was performed with a feedback rate of 1 Hz while monotonically translating the sample stage over the $120\text{ }\mu\text{m}$ range on mouse skull exposed to air and that immersed beneath a $500\text{ }\mu\text{m}$ -thick liquid layer, respectively. A total of 240 effective frames were acquired, equivalent to an axial motion rate of $2\text{ }\mu\text{m/s}$. Observation of the skull surface was permitted throughout the recording session for both experiments in [Visualization 3](#) and [Visualization 5](#), despite with slight deviations that may have been caused by rapid vibrations from manual stage translation. In comparison, [Visualization 4](#) and [Visualization 6](#) only recorded structural features coming in and out of focus when the natural focal plane was within close proximity of the skull surface. In order to quantitatively analyse the imaging performance with and without autofocusing, Pearson correlation coefficients were computed for each image frame with respect to that acquired at $0\text{ }\mu\text{m}$ defocus position. Results for [Visualization 3](#) and [Visualization 4](#) are given as example. Figure 6(f) plots correlation coefficients as a function of the defocus position. The yellow curve depicts imaging without autofocusing, showing strong correlation only within a narrow region about the reference plane, before dropping rapidly to below 0.2 [39]. The case with autofocusing on is illustrated in blue, where troughs represent ‘blinking’ patterns during shift measurements. Red asterisks mark frames before (after) seeing a sudden decrease (increase) in correlation, which contain those captured in between shift measurements, resulting in values above 0.4 . This falls within expectations considering that the focal plane may not be static during the course of acquiring one image and lateral shifts may occur during manual stage translation. Figure 6(g) sorts the coefficients in descending order and plots the number of frames above a certain value. Dotted lines indicate a strong (>0.7) or moderate (>0.5) relationship between the reference image [39], and a significant increase can be seen when performing regular tracking routines.

In the final experiment, the experimenter’s finger was placed on a fixed mount below the objective lens while using the other hand for focus control. In [Visualization 7](#) with autofocusing, the microscope retained focus throughout the whole recording session; whereas in [Visualization 8](#) without autofocusing, the focal plane constantly moved with movements of the finger during breathing, such that it was difficult to concentrate on any specific feature.

4. Discussion

The proposed autofocusing technique is unique in a number of ways. First, we directly take phase measurements with a dedicated wavefront sensor instead of interpreting defocus information from sample images, which prevents the autofocusing range from being limited by the availability of structural details. A $120\text{ }\mu\text{m}$ range was achieved, and could be further extended by using a SHWS with larger dynamic range. In addition, the SHWS permitted simultaneous correction of system aberrations during RF by incorporating this information within calibrated DM control voltages. Therefore, any operation that involved the use of RF voltages also corrected for system aberrations at each axial location, including prior to making phase measurements when acquiring training

datasets, the ground truth data itself that constituted of RF voltages, as well as prior to making each shift measurement in real time. Other aberrations induced by non-ideal surface conditions and refractive index mismatch between different media could also be detected. Acquisition of training data under these diverse conditions greatly enhanced the trained network's generality towards non-ideal scenarios and allowed accurate repositioning of the focal plane. Moreover, as the SHWS prevented network exposure to sample structures, a second aspect of generality, that towards different specimen types, was simultaneously permitted. In contrast to implementing the proposed technique in a low NA system similar to that used in this work, high NA systems would greatly benefit from simultaneous correction of aberrations induced by refractive index mismatch and sample inhomogeneity [40] by using a similar adaptive optics correction routine to that set out in Section 2.2 at predetermined step intervals during training data acquisition. Second, the implementation of RF allowed not only the correction of system aberrations, but also axial refocusing to be performed in less than 1 ms, which is much faster than movements of the sample stage and objective lens. This is beneficial in particular for surgical applications to stabilise imaging during respiratory and pulsing movements over long distances. Third, compared to CNN-based approaches, the use of a Bi-LSTM network for learning of sequential aberration information from a SHWS is more tailored to the task at hand. By removing irrelevant information, such as cell morphology and size etc., from the network input, more concentration can be placed on factors that determine the position and quality of the focal spot. Consequently, the network is not only simpler and faster to train, but also leads to unique strengths such as allowing for higher generality and image quality.

It is worth mentioning that the autofocusing range depends on the NA of the objective lens, the mechanical stroke of the DM, the signal-to-noise ratio (SNR) at the SHWS, as well as the dynamic range of the SHWS. The first two factors determine the microscope's maximum permissible RF range, while the latter two determine the maximum detectable phase aberration. A lower NA, longer mechanical stroke, higher SNR, and larger dynamic range potentially lead to a more extended defocus range. Although the SNR wasn't of special concern in this system for the reported defocus range and exposure time, it can decrease at larger distances away from the specimen surface and may become a limiting factor for systems with low photon counts. A longer SHWS exposure time can then be adopted, but at the expense of speed, which may nevertheless be favourable for applications that do not require autofocusing in real time. The maximum FOV over which the network trained with small FOVs would remain stable is dictated by the isoplanatic patch [41], within which the SHWS is able to obtain relatively constant phase measurements. At superficial layers of the specimen, its size generally decreases with a higher NA. The downside of this can be alleviated by performing imaging and phase aberration measurements alternately, such that autofocusing is performed over a small FOV (backscattered light) or at a fixed lateral position (fluorescence), while imaging is performed over much larger FOVs. This also enhances autofocusing speed by lowering the SHWS exposure time for phase measurements. Speed is then only ultimately limited by the DM settling time. For the current system, however, it was only possible to perform imaging in parallel with phase aberration measurements, which compromised the autofocusing speed in reality. Despite so, the alternation of these two processes would be a task for the future. Finally, we would like to point out that compared to existing single-shot methods, such as those reported in [7,13,17], although the requirement for three measurements makes the proposed method comparably slower, the additional merits of allowing for a much larger autofocusing range, and system-aberration-free imaging over the full extended range compensates for the minor decrease in speed.

5. Conclusion

In summary, we demonstrate a new autofocusing method based on RF and sequence-dependent learning. By taking advantage of a DM and SHWS, aberration-free autofocusing was achieved in

tissue both exposed to air and immersed in liquid. An extended 120 μm range was obtained with a positioning error smaller than half the axial resolution. The technique is valid for FOVs of different sizes and for specimens not seen in the network training process. Continuous tracking of sample motion was enabled in real time on a compact neurosurgical microscope, showing great potential for transition towards clinical and surgical applications. Finally, we note that although this work was demonstrated using reflectance contrast, the technique is fully applicable to fluorescence imaging by incorporating a band-pass or short-pass filter before the SHWS to detect back-scattered illumination light.

Funding. European Research Council (695140, 812998).

Acknowledgments. We thank Carla Schmidt for help with sample preparation.

Disclosures. The authors declare no conflicts of interest.

Data availability. There are 8 Visualizations associated with this paper. If they cannot be directly accessed through links in this paper, please contact the corresponding author. Other data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

Supplemental document. See [Supplement 1](#) for supporting content.

References

1. M. Montalto, R. McKay, and R. Filkins, "Autofocus methods of whole slide imaging systems and the introduction of a second-generation independent dual sensor scanning method," *J. Pathol. Inform.* **2**(1), 44 (2011).
2. Z. Bian, C. Guo, S. Jiang, J. Zhu, R. Wang, P. Song, Z. Zhang, K. Hoshino, and G. Zheng, "Autofocusing technologies for whole slide imaging and automated microscopy," *J. Biophotonics* **13**(12), e202000227 (2020).
3. P. Langehanenberg, G. von Bally, and B. Kemper, "Autofocusing in digital holographic microscopy," *3D Res.* **2**(1), 4 (2011).
4. L. Ma and B. Fei, "Comprehensive review of surgical microscopes: technology development and medical applications," *J. Biomed. Opt.* **26**(01), 010901 (2021).
5. R. Redondo, G. Cristóbal, G. B. Garcia, O. Deniz, J. Salido, M. del Milagro Fernandez, J. Vidal, J. C. Valdiviezo, R. Nava, B. Escalante-Ramírez, and M. Garcia-Rojo, "Autofocus evaluation for brightfield microscopy pathology," *J. Biomed. Opt.* **17**(3), 036008 (2012).
6. S. Yazdanfar, K. B. Kenny, K. Tasimi, A. D. Corwin, E. L. Dixon, and R. J. Filkins, "Simple and robust image-based autofocusing for digital microscopy," *Opt. Express* **16**(12), 8670–8677 (2008).
7. J. Liao, L. Bian, Z. Bian, Z. Zhang, C. Patel, K. Hoshino, Y. C. Eldar, and G. Zheng, "Single-frame rapid autofocusing for brightfield and fluorescence whole slide imaging," *Biomed. Opt. Express* **7**(11), 4763–4768 (2016).
8. J. Liao, Z. Wang, Z. Zhang, Z. Bian, K. Guo, A. Nambiar, Y. Jiang, S. Jiang, J. Zhong, M. Choma, and G. Zheng, "Dual light-emitting diode-based multichannel microscopy for whole-slide multiplane, multispectral and phase imaging," *J. Biophotonics* **11**(2), e201700075 (2018).
9. J. Xu, X. Tian, X. Meng, Y. Kong, S. Gao, H. Cui, F. Liu, L. Xue, C. Liu, and S. Wang, "Wavefront-sensing-based autofocusing in microscopy," *J. Biomed. Opt.* **22**(8), 086012 (2017).
10. J. Xu, Y. Kong, Z. Jiang, S. Gao, L. Xue, F. Liu, C. Liu, and S. Wang, "Accelerating wavefront-sensing-based autofocusing using pixel reduction in spatial and frequency domains," *Appl. Opt.* **58**(11), 3003–3012 (2019).
11. L. Wei and E. Roberts, "Neural network control of focal position during time-lapse microscopy of cells," *Sci. Rep.* **8**(1), 7313 (2018).
12. S. Jiang, J. Liao, Z. Bian, K. Guo, Y. Zhang, and G. Zheng, "Transform- and multi-domain deep learning for single-frame rapid autofocusing in whole slide imaging," *Biomed. Opt. Express* **9**(4), 1601–1612 (2018).
13. Z. Ren, Z. Xu, and E. Y. Lam, "Learning-based nonparametric autofocusing for digital holography," *Optica* **5**(4), 337–344 (2018).
14. Y. Wu, Y. Rivenson, Y. Zhang, Z. Wei, H. Günaydin, X. Lin, and A. Ozcan, "Extended depth-of-field in holographic imaging using deep-learning-based autofocusing and phase recovery," *Optica* **5**(6), 704–710 (2018).
15. Y. Wu, Y. Rivenson, H. Wang, Y. Luo, E. Ben-David, L. A. Bentolila, C. Pritz, and A. Ozcan, "Three-dimensional virtual refocusing of fluorescence microscopy images using deep learning," *Nat. Methods* **16**(12), 1323–1331 (2019).
16. H. Pinkard, Z. Phillips, A. Babakhani, D. A. Fletcher, and L. Waller, "Deep learning for single-shot autofocus microscopy," *Optica* **6**(6), 794–797 (2019).
17. Y. Luo, L. Huang, Y. Rivenson, and A. Ozcan, "Single-shot autofocusing of microscopy images using deep learning," *ACS Photonics* **8**(2), 625–638 (2021).
18. Q. Li, X. Liu, J. Jiang, C. Guo, X. Ji, and X. Wu, "Rapid whole slide imaging via dual-shot deep autofocusing," *IEEE Trans. Comput. Imaging* **7**, 124–136 (2021).
19. C. Belthangady and L. A. Royer, "Applications, promises, and pitfalls of deep learning for fluorescence image reconstruction," *Nat. Methods* **16**(12), 1215–1225 (2019).
20. B. C. Platt and R. Shack, "History and principles of shack-hartmann wavefront sensing," *J. Refract. Surg.* **17**(5), S573–S577 (2001).

21. M. Born and E. Wolf, *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light* (Elsevier, 2013).
22. M. J. Booth, "Adaptive optical microscopy: the ongoing quest for a perfect image," *Light: Sci. Appl.* **3**(4), e165 (2014).
23. F. Zernike, "Diffraction theory of the knife-edge test and its improved form, the phase-contrast method," *Mon. Not. R. Astron. Soc.* **94**(5), 377–384 (1934).
24. V. N. Mahajan, "Zernike circle polynomials and optical aberrations of systems with circular pupils," *Appl. Opt.* **33**(34), 8121–8124 (1994).
25. S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.* **9**(8), 1735–1780 (1997).
26. M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.* **45**(11), 2673–2681 (1997).
27. G. D. Reddy, K. Kelleher, R. Fink, and P. Saggau, "Three-dimensional random access multiphoton microscopy for functional imaging of neuronal activity," *Nat. Neurosci.* **11**(6), 713–720 (2008).
28. B. F. Grewe, D. Langer, H. Kasper, B. M. Kampa, and F. Helmchen, "High-speed in vivo calcium imaging reveals neuronal network activity with near-millisecond precision," *Nat. Methods* **7**(5), 399–405 (2010).
29. B. F. Grewe, F. F. Voigt, M. van't Hoff, and F. Helmchen, "Fast two-layer two-photon imaging of neuronal cell populations using an electrically tunable lens," *Biomed. Opt. Express* **2**(7), 2035–2046 (2011).
30. E. J. Botcherby, R. Juškaitis, M. J. Booth, and T. Wilson, "An optical technique for remote focusing in microscopy," *Opt. Commun.* **281**(4), 880–887 (2008).
31. M. Dal Maschio, A. M. De Stasi, F. Benfenati, and T. Fellin, "Three-dimensional in vivo scanning microscopy with inertia-free focus control," *Opt. Lett.* **36**(17), 3503–3505 (2011).
32. Y. Yang, W. Chen, J. L. Fan, and N. Ji, "Adaptive optics enables aberration-free single-objective remote focusing for two-photon fluorescence microscopy," *Biomed. Opt. Express* **12**(1), 354–366 (2021).
33. J. Cui, R. Turcotte, K. M. Hampson, M. Wincott, C. C. Schmidt, N. J. Emptage, P. Charalampaki, and M. J. Booth, "Compact and contactless reflectance confocal microscope for neurosurgery," *Biomed. Opt. Express* **11**(8), 4772–4785 (2020).
34. M. Rueckel, J. A. Mack-Bucher, and W. Denk, "Adaptive wavefront correction in two-photon microscopy using coherence-gated wavefront sensing," *Proc. Natl. Acad. Sci.* **103**(46), 17137–17142 (2006).
35. X. Tao, B. Fernandez, O. Azucena, M. Fu, D. Garcia, Y. Zuo, D. C. Chen, and J. Kubby, "Adaptive optics confocal microscopy using direct wavefront sensing," *Opt. Lett.* **36**(7), 1062–1064 (2011).
36. K. Wang, W. Sun, C. T. Richie, B. K. Harvey, E. Betzig, and N. Ji, "Direct wavefront sensing for high-resolution in vivo imaging in scattering tissue," *Nat. Commun.* **6**(1), 7276 (2015).
37. V. N. Mahajan, "Strehl ratio for primary aberrations: some analytical results for circular and annular pupils," *J. Opt. Soc. Am.* **72**(9), 1258–1266 (1982).
38. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980 (2014).
39. D. S. Moore and S. Kirkland, *The basic practice of statistics*, vol. 2 (WH Freeman, 2007).
40. L. Silvestri, M. C. Müllenbroich, I. Costantini, A. P. Di Giovanna, G. Mazzamuto, A. Franceschini, D. Kutra, A. Kreshuk, C. Checcucci, L. O. Toresano, P. Frasconi, L. Sacconi, and F. S. Pavone, "Universal autofocus for quantitative volumetric microscopy of whole mouse brains," *Nat. Methods* **18**(8), 953–958 (2021).
41. D. L. Fried, "Anisoplanatism in adaptive optics," *J. Opt. Soc. Am.* **72**(1), 52–61 (1982).