optica

# Deep learning for single-shot autofocus microscopy

Henry Pinkard,[1,2,3,4,*] Zachary Phillips,[5] Arman Babakhani,[6] Daniel A. Fletcher,[7,8] and Laura Waller[2,3,5,8]

[1] *Computational Biology Graduate Group, University of California, Berkeley, California 94720, USA*
[2] *Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, California 94720, USA*
[3] *Berkeley Institute for Data Science, Berkeley, California 94720, USA*
[4] *University of California San Francisco Bakar Computational Health Sciences Institute, San Francisco, California 94158, USA*
[5] *Graduate Group in Applied Science and Technology, University of California, Berkeley, California 94720, USA*
[6] *Department of Physics, University of California, Berkeley, California 94720, USA*
[7] *Department of Bioengineering and Biophysics Program, University of California, Berkeley, California 94720, USA*
[8] *Chan Zuckerberg Biohub, San Francisco, California 94158, USA*
**Corresponding author: henry.pinkard@gmail.com*

*Received 14 March 2019; revised 12 May 2019; accepted 15 May 2019 (Doc. ID 362084); published 5 June 2019*

**Maintaining an in-focus image over long time scales is an essential and nontrivial task for a variety of microscopy applications. Here, we describe a fast, robust autofocusing method compatible with a wide range of existing microscopes. It requires only the addition of one or a few off-axis illumination sources (e.g., LEDs), and can predict the focus correction from a single image with this illumination. We designed a neural network architecture, the fully connected Fourier neural network (FCFNN), that exploits an understanding of the physics of the illumination to make accurate predictions with 2–3 orders of magnitude fewer learned parameters and less memory usage than existing state-of-the-art architectures, allowing it to be trained without any specialized hardware. We provide an open-source implementation of our method, to enable fast, inexpensive autofocus compatible with a variety of microscopes.**

https://doi.org/10.1364/OPTICA.6.000794

Many biological experiments involve imaging samples in a microscope over long time periods or large spatial scales, making it difficult to keep the sample in focus. When observing a sample over time periods of hours or days, for example, thermal fluctuations can induce focus drift [1]. Or, when scanning and stitching together many fields-of-view (FoV) to form a high-content, high-resolution image, a sample that is not sufficiently flat necessitates refocusing at each position [2]. Since it is often experimentally impractical or cumbersome to manually maintain focus, an automatic focusing mechanism is essential.

A variety of solutions have been developed for autofocus. Broadly, these methods can be divided into two classes: hardware-based schemes that attempt to directly measure the distance from the objective lens to the sample [3–7], and software-based methods that take one or more out-of-focus images and use them to determine the optimal focal position [8–11]. The former usually require hardware modifications to the microscope (e.g., an infrared laser interferometry setup, additional cameras, or optical elements), which can be expensive and place constraints on other aspects of the imaging system. Software-based methods, on the other hand, can be slow or inaccurate. A software-based method, for example, might require a full focal stack, then use some measure of image sharpness to compute the ideal focal plane [8]. More advanced methods attempt to reduce the number of images needed to compute the correct focus [9], or use just a single out-of-focus image [10,11]. However, existing single-shot autofocus methods either rely on nontrivial hardware modifications such additional lenses and sensors [11] or are limited in their application to specialized regimes (i.e., can only correct defocus in one direction within a certain range) [10].

Here, we demonstrate a new computational imaging-based, single-shot autofocus method that does not suffer from the limitations of previous methods. The only hardware modification it requires is the addition of one or more off-axis LEDs as an illumination source, from which we correct defocus based on a single out-of-focus image. Alternately, it can be used with no hardware modification on existing coded-illumination setups, which have been demonstrated for super-resolution [12–14], quantitative phase [12,13,15], and multicontrast microscopy [16,17].

The central idea of our method is that a neural network can be trained to predict how far out of focus a microscope is, based on a single image taken at an arbitrary defocus under spatially coherent illumination. A related idea has recently been used to achieve fast, post-experimental digital refocusing in digital holography [18,19]. Our work addresses autofocusing in more general microscope systems, with both incoherent and coherent illumination. Intuitively, we believe this works because coherent illumination yields images with sharp features even when the sample is out of focus. Thus, there is sufficient information in the out-of-focus image that an appropriate neural network can learn a function that maps these features to the correct defocus distance, regardless of the structural details of the sample. To test this idea we collected data using a ZEISS Axio Observer microscope

2334-2536/19/060794-04 Journal © 2019 Optical Society of America

(20×, 0.5 NA) with the illumination source replaced by a programmable quasi-dome LED array [20]. The LED array provides a flexible means of source patterning, but is not necessary to implement this technique (see Note S1).

Though our experimental focus prediction requires only one image, we do need to collect focal stacks for training and validation. We use Micro-Magellan [21] for software control of the microscope, collecting focal stacks over 60 μm with 1 μm spacing, distributed symmetrically around the true focal plane. For each part of the sample, we collect focal stacks with two different types of illumination: spatially coherent (i.e., a single LED) and (nearly) spatially incoherent (i.e., many LEDs at once).

The incoherent focal stack is used for computing the ground truth focal position, since the reduced coherence results in sharp images only when the sample is in focus. Sharpness can be quantified for each image in the stack by summing the high-frequency content of its radially averaged log power spectrum. The maximum of the resultant curve was chosen as the ground truth focal position for the stack [Fig. 1(a), left]. Because this ground truth value is calculated by a deterministic algorithm, this paradigm scales well to large amounts of training data. For transparent samples, the incoherent image stack was captured with asymmetric illumination to create phase contrast [22]. In our case, this was achieved by using the LED array to project a half annulus source pattern [15]; however, any asymmetric source pattern should suffice.

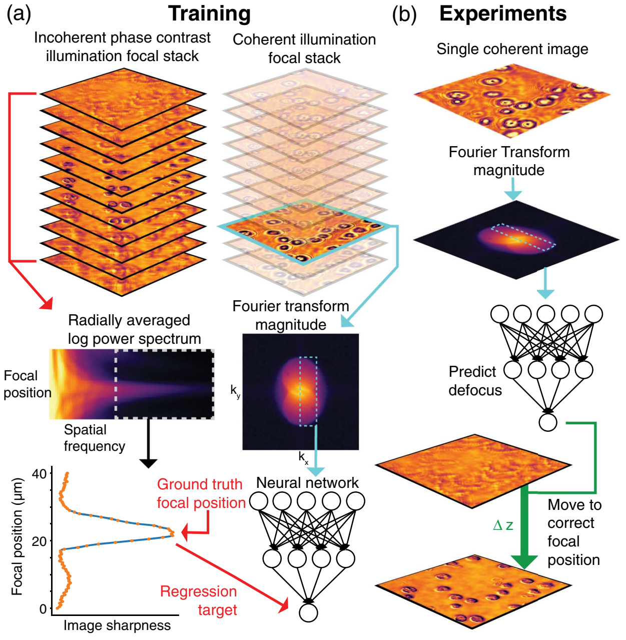
(a)
Training
Incoherent phase contrast illumination focal stack
Coherent illumination focal stack
(b)
Experiments
Single coherent image
Fourier Transform magnitude
Radially averaged log power spectrum
Fourier transform magnitude
Focal position
Spatial frequency
$k_y$
$k_x$
Predict defocus
40
20
0
Focal position (μm)
Image sharpness
Ground truth focal position
Neural network
Regression target
Δz
Move to correct focal position

**Fig. 1.** Training and defocus prediction. (a) Training data consists of two focal stacks for each part of the sample, one with incoherent (phase contrast) illumination, and one with off-axis coherent illumination. Left: The high spatial frequency part of each image's power spectrum from the incoherent stack is used to compute a ground truth focal position. Right: For each coherent image in the stack, the central pixels from the magnitude of its Fourier transform are used as input to a neural network trained to predict defocus. The full set of training examples is generated by repeating this process for each of the coherent images in the stack. (b) After training, experiments need only collect a single coherent image, which is fed through the same pipeline to predict defocus. The microscope's focus can then be adjusted to correct defocus.

The coherent focal stack is used one image at a time as the input to the network, which is trained to predict the ground truth focal position (Fig. 1). Since the network only takes a single image as its input, each image in the stack represents a separate training example. In our case, the coherent focal stack was captured by illuminating the sample with a single off-axis LED. In the case of arbitrary illumination control (e.g., with an LED array) different illumination angles or patterns may perform differently for a given amount of training data. Supplementary Fig. S1 compares performance for varying single-LED illumination angles as well as multi-LED patterns. For simplicity, here we consider only the case of a single LED positioned at an angle of 24 deg relative to the optical axis.

Our neural network architecture for predicting defocus (described in detail in Note S3), which we call the fully connected Fourier neural network (FCFNN), differs substantially from the convolutional neural networks (CNNs) typically used in image processing tasks [18,19,23] (Note S4). We reasoned that singly scattered light would contain the most useful information for defocus prediction, and thus we designed the FCFNN to exclude parts of the captured image's Fourier transform that are outside the single-scattering region for off-axis illumination (Fig. S2). This results in 2–3 orders of magnitude fewer free parameters and memory usage during training than state-of-the-art CNNs (Table S1). Hence, our network can be trained on a desktop CPU in a few hours with no specialized computing hardware, which we believe makes our method more reproducible, without sacrificing quality.

Briefly, the FCFNN [Fig. 1(a), right] begins with a single coherent image. This image is Fourier transformed, and the magnitude of the complex-valued pixels in the central part of the Fourier transform are reshaped into a single vector, which is used as the input to a trainable fully connected neural network. After the network has been trained, it can be used to correct defocus during an experiment by capturing a single image at an arbitrary defocus under the same coherent illumination. The network predicts defocus distance, then the microscope moves to the correct focal position [Fig. 1(b)].

Training with 440 focal stacks took 1.5 h on a desktop CPU or 30 min on a GeForce GTX 1080 Ti GPU, in addition to 2 min per focal stack for pre-computing ground truth focal planes and Fourier transforms. A single prediction from a 2048 × 2048 image takes ~50 ms on a desktop CPU. We were able to train FCFNNs capable of predicting defocus with root-mean-squared error (RMSE) smaller than the axial thickness of the sample (cells). Figure 2 shows how this performance varies based on the number of focal stacks used to train the network, where each focal stack contained 60 planes spaced 1 μm apart, distributed symmetrically around the true focal plane. Note that this curve could be quite different depending on the sample type and quality of training data.

To test the performance of our method across different samples, we collected data from two different sample types [Fig. 3(a)]: white blood cells attached to coverglass, and an unstained 5 μm thick mounted histology tissue section. When the network is *trained* on images of cells, then *tested* on different images of cells, it performs very well [Fig. 3(b)]. However, when the network is trained on images of cells, then tested on a different sample type (tissue), it performs poorly [Fig. 3(c)]. Hence, the method does not inherently generalize to new sample types. To solve this

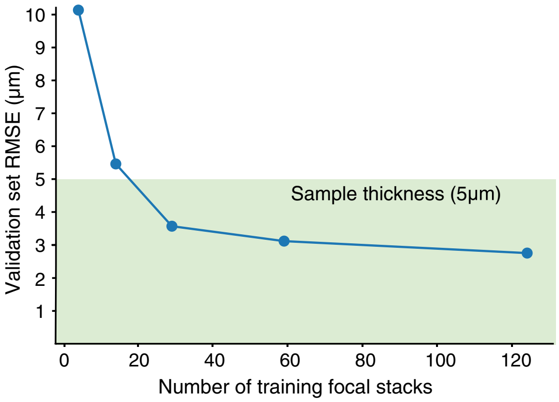

**Fig. 2.** Performance versus amount of training data. Defocus prediction performance (measured by validation RMSE) improves as a function of the number of focal stacks used during the training phase of the method.

problem, we diversify the training data. We add a smaller amount of additional training data from the new sample type (in this case, 130 focal stacks of tissue data, in addition to the 440 stacks of cell data it was originally trained on). With this training, the network performs well on both tissue and cell samples. Hence, our method can generalize to other sample types, without sacrificing performance on the original sample type [Fig. 3(d)]. The best performing neural networks in other domains are typically trained on large and varied datasets [24]. Thus, if the FCFNN is trained on defocus data from a variety of sample types, it should generalize to new types more easily.

Empirically, we discovered that discarding the phase of the Fourier transform and using only the magnitude as the input to the network dramatically boosted performance. To illustrate, Fig. 4(a) compares networks trained using the Fourier transform magnitude as the input versus those trained on the argument of the Fourier transform phase. Not only were networks using magnitude able to better fit the training data, they also generalized better to a validation set. This suggests useful information for predicting defocus in a coherent intensity image is relatively more concentrated in the magnitude compared to the phase of its Fourier transform. We speculate that this happens because the phase of the intensity image generally relates more to the spatial position of features (which is unimportant for focus prediction), whereas the magnitude contains more information about how they are transformed by the imaging system.

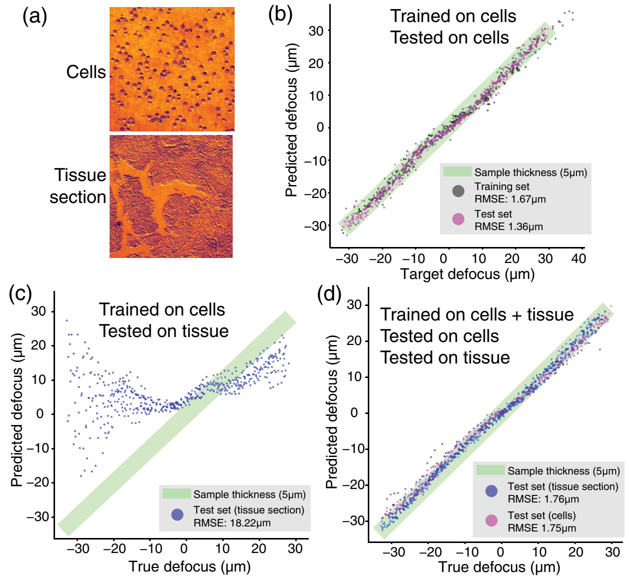

**Fig. 3.** Generalization to new sample types. (a) Representative images of cells and tissue section samples. (b) A network trained on focal stacks of cells predicts defocus well in other cell samples. (c) This network, however, fails when predicting defocus in tissue sections. (d) After adding limited additional training data on tissue section samples, however, the network can learn to predict defocus well in both sample types.

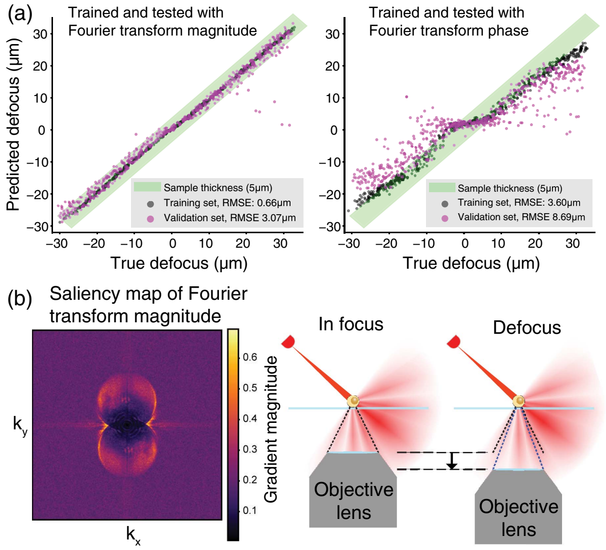

**Fig. 4.** Understanding how the network predicts defocus. (a) A network trained on the magnitude of the Fourier transform of the input image performs better than one trained on the argument of the phase of the Fourier transform. (b) Left: A saliency map (the magnitude of the defocus prediction's gradient with respect to the Fourier transform magnitude) shows the edges of the object spectrum have the strongest influence on defocus predictions. Right: Edges correspond to high-angle scattered light, which may not be captured off-focus, providing significant changes in the input image with defocus.

To understand what features of the images the network learns to use to make predictions, we compute a saliency map for a network trained using the entire uncropped Fourier transform, shown in Fig. 4(b). The saliency map attempts to identify which parts of the input the network is using to make decisions, by visualizing the gradient of a single unit within the neural network with respect to the input [25]. The idea is that the output unit is more sensitive to features with a large gradient and thus these have a greater influence on prediction. In our case, the gradient of the output (i.e., the defocus prediction) was computed with respect to the Fourier transform magnitude. Averaging the magnitude of the gradient image over many examples clearly shows that the network recognizes specific parts of the overlapping two-circle structure [Fig. 4(b)] that is typical for an image formed by coherent off-axis illumination (Fig. S2) [26]. In particular, the regions at the edges of the circles have an especially large gradient. These areas correspond to the highest angles of light collected by the objective lens. Intuitively, this makes sense because changing the focus will lead to proportionally greater changes in the light collected at the highest angles [Fig. 4(b)].

To summarize, we have demonstrated a method to train and use neural networks for single-shot autofocus, with an analysis of design principles and practical trade-offs. The method works with different sample types and is simple to implement on a conventional transmitted light microscope, requiring only the addition of off-axis illumination and no specialized hardware for training the neural network. We introduced the FCFNN, a neural network architecture that incorporates knowledge of the physics of the imaging system into its design, thereby making it orders of magnitude more efficient in terms of the parameter number and memory requirements during training than general state-of-the-art approaches for image processing.

The Code 1, Ref. [27] needed to implement this technique and reproduce all figures in this Letter can be found in the Jupyter notebook. Due to its large size, the corresponding data is available upon request.

**Funding.** National Science Foundation (NSF) (DMR 1548924); National Institutes of Health (NIH) (R01); Gordon and Betty Moore Foundation (GBMF3834); Alfred P. Sloan Foundation (2013-10-27); David and Lucile Packard Foundation; Chan Zuckerberg Biohub; Koret Foundation.

**Acknowledgment.** The authors thank S. Y. Liu, for providing the tissue sample, and BIDS and its personnel for providing physical space and general logistical and technical support. This project was funded by Packard Fellowship and Chan Zuckerberg Biohub Investigator Awards to Laura Waller and Daniel Fletcher; STROBE: A NSF Science & Technology Center; an NIH grant to Daniel Fletcher; an NSF Graduate Research Fellowship awarded to Henry Pinkard; and a Berkeley Institute for Data Science/UCSF Bakar Computational Health Sciences Institute Fellowship awarded to Henry Pinkard with support from the Koret Foundation, the Gordon and Betty Moore Foundation, and the Alfred P. Sloan Foundation to the University of California, Berkeley.

See Supplement 1 for supporting content.

## REFERENCES

1. M. Kreft, M. Stenovec, and R. Zorec, Ann. N.Y. Acad. Sci. **1048**, 321 (2005).
2. M. D. Zarella, D. Bowman, F. Aeffner, N. Farahani, A. Xthona, S. F. Absar, A. Parwani, M. Bui, and D. J. Hartman, Arch. Pathol. Lab. Med. **143**, 222 (2019).
3. Nikon Instruments Inc., "Nikon perfect focus," 2019, https://www.microscopyu.com/applications/live-cell-imaging/nikon-perfect-focus-system.
4. ZEISS, "Zeiss definite focus," 2019, https://www.zeiss.com/microscopy/us/products/light-microscopes/axio-observer-for-biology/definite-focus.html.
5. K. Guo, J. Liao, Z. Bian, X. Heng, and G. Zheng, Biomed. Opt. Express **6**, 3210 (2015).
6. M. Bathe-Peters, P. Annibale, and M. J. Lohse, Opt. Express **26**, 2359 (2018).
7. X. Zhang, F. Zeng, Y. Li, and Y. Qiao, Opt. Express **26**, 887 (2018).
8. F. Shen, L. Hodgson, and K. Hahn, "Digital autofocus methods for automated microscopy," in *Methods in Enzymology* (Academic, 2006), Vol. **414**, pp. 620–632.
9. S. Yazdanfar, K. B. Kenny, K. Tasimi, A. D. Corwin, E. L. Dixon, and R. J. Filkins, Opt. Express **16**, 8670 (2008).
10. J. Liao, Y. Jiang, Z. Bian, B. Mahrou, A. Nambiar, A. W. Magsam, K. Guo, S. Wang, Y. K. Cho, and G. Zheng, Opt. Lett. **42**, 3379 (2017).
11. J. Liao, L. Bian, Z. Bian, Z. Zhang, C. Patel, K. Hoshino, Y. C. Eldar, and G. Zheng, Biomed. Opt. Express **7**, 4763 (2016).
12. G. Zheng, R. Horstmeyer, C. Yang, G. Zheng, and C. Yang, Nat. Photonics **7**, 739 (2013).
13. X. Ou, R. Horstmeyer, C. Yang, and G. Zheng, Opt. Lett. **38**, 4845 (2013).
14. L. Tian and L. Waller, Optica **2**, 104 (2015).
15. L. Tian and L. Waller, Opt. Express **23**, 11394 (2015).
16. G. Zheng, C. Kolner, and C. Yang, Opt. Lett. **36**, 3987 (2011).
17. Z. Liu, L. Tian, S. Liu, and L. Waller, J. Biomed. Opt. **19**, 106002 (2014).
18. Y. Wu, Y. Rivenson, Y. Zhang, Z. Wei, H. Günaydin, X. Lin, and A. Ozcan, Optica **5**, 704 (2018).
19. Z. Ren, Z. Xu, and E. Y. Lam, Optica **5**, 337 (2018).
20. Z. F. Phillips, R. Eckert, and L. Waller, "Quasi-dome: a self-calibrated high-NA LED illuminator for Fourier ptychography," in *Imaging and Applied Optics (3D, AIO, COSI, IS, MATH, pcAOP)* (2017), Vol. IW4E.5.
21. H. Pinkard, N. Stuurman, K. Corbin, R. Vale, and M. F. Krummel, Nat. Methods **13**, 807 (2016).
22. S. B. Mehta and C. J. R. Sheppard, Opt. Lett. **34**, 1924 (2009).
23. S. J. Yang, M. Berndl, D. M. Ando, M. Barch, A. Narayanaswamy, E. Christiansen, S. Hoyer, C. Roat, J. Hung, C. T. Rueden, A. Shankar, S. Finkbeiner, and P. Nelson, BMC Bioinf. **19**, 28 (2018).
24. A. Krizhevsky, I. Sutskever, and G. E. Hinton, *Advances in Neural Information Processing Systems* (2011), pp. 1097–1105.
25. K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: visualising image classification models and saliency maps," arXiv:1312.6034 (2013).
26. R. Eckert, Z. F. Phillips, and L. Waller, Appl. Opt. **57**, 5434 (2018).
27. H. Pinkard, "Single-shot autofocus microscopy using deep learning–code," (2019), https://doi.org/10.6084/m9.figshare.7453436.v1.