Research Article | Vol. 9, No. 4 | 1 Apr 2018 | BIOMEDICAL OPTICS EXPRESS 1601
Biomedical Optics EXPRESS

Transform- and multi-domain deep learning for single-frame rapid autofocusing in whole slide imaging

SHAOWEI JIANG,[1,4] JUN LIAO,[1,4] ZICHAO BIAN,[1] KAIKAI GUO,[1] YONGBING ZHANG,[2] AND GUOAN ZHENG[1,3,*]

[1]*Biomedical Engineering, University of Connecticut, Storrs, CT, 06269, USA*
[2]*Shenzhen Key Lab of Broadband Network and Multimedia, Graduate School at Shenzhen, Tsinghua University, Shenzhen, 518055, China*
[3]*Electrical and Computer Engineering, University of Connecticut, Storrs, CT, 06269, USA*
[4]*These authors contributed equally to this work*
guoan.zheng@uconn.edu

Abstract: A whole slide imaging (WSI) system has recently been approved for primary diagnostic use in the US. The image quality and system throughput of WSI is largely determined by the autofocusing process. Traditional approaches acquire multiple images along the optical axis and maximize a figure of merit for autofocusing. Here we explore the use of deep convolution neural networks (CNNs) to predict the focal position of the acquired image without axial scanning. We investigate the autofocusing performance with three illumination settings: incoherent Kohler illumination, partially coherent illumination with two plane waves, and one-plane-wave illumination. We acquire ~130,000 images with different defocus distances as the training data set. Different defocus distances lead to different spatial features of the captured images. However, solely relying on the spatial information leads to a relatively bad performance of the autofocusing process. It is better to extract defocus features from transform domains of the acquired image. For incoherent illumination, the Fourier cutoff frequency is directly related to the defocus distance. Similarly, autocorrelation peaks are directly related to the defocus distance for two-plane-wave illumination. In our implementation, we use the spatial image, the Fourier spectrum, the autocorrelation of the spatial image, and combinations thereof as the inputs for the CNNs. We show that the information from the transform domains can improve the performance and robustness of the autofocusing process. The resulting focusing error is ~0.5 μm, which is within the 0.8-μm depth-of-field range. The reported approach requires little hardware modification for conventional WSI systems and the images can be captured on the fly without focus map surveying. It may find applications in WSI and time-lapse microscopy. The transform- and multi-domain approaches may also provide new insights for developing microscopy-related deep-learning networks. We have made our training and testing data set (~12 GB) open-source for the broad research community.

OCIS codes: (180.5810) Scanning microscopy; (170.4730) Optical pathology; (100.4996) Pattern recognition, neural networks

References and links

1. S. Al-Janabi, A. Huisman, and P. J. Van Diest, "Digital pathology: current status and future perspectives," Histopathology **61**(1), 1–9 (2012).
2. E. Abels and L. Pantanowitz, "Current state of the regulatory trajectory for whole slide imaging devices in the USA," J. Pathol. Inform. **8**(1), 23 (2017).
3. J. Liao, Y. Jiang, Z. Bian, B. Mahrou, A. Nambiar, A. W. Magsam, K. Guo, S. Wang, Y. K. Cho, and G. Zheng, "Rapid focus map surveying for whole slide imaging with continuous sample motion," Opt. Lett. **42**(17), 3379–3382 (2017).

#319811 https://doi.org/10.1364/BOE.9.001601
Journal © 2018 Received 18 Jan 2018; revised 1 Mar 2018; accepted 2 Mar 2018; published 8 Mar 2018

Biomedical Optics EXPRESS

4. M. C. Montalto, R. R. McKay, and R. J. 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).
5. 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).
6. K. Guo, J. Liao, Z. Bian, X. Heng, and G. Zheng, "InstantScope: a low-cost whole slide imaging system with instant focal plane detection," Biomed. Opt. Express **6**(9), 3210–3216 (2015).
7. L. Yu, H. Chen, Q. Dou, J. Qin, and P.-A. Heng, "Automated melanoma recognition in dermoscopy images via very deep residual networks," IEEE Trans. Med. Imaging **36**(4), 994–1004 (2017).
8. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016), 770–778.
9. J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016), 1646–1654.
10. C. Feichtenhofer, A. Pinz, and R. Wildes, "Spatiotemporal residual networks for video action recognition," in *Advances in Neural Information Processing Systems*, 2016), 3468–3476.
11. P. Langehanenberg, G. von Bally, and B. Kemper, "Autofocusing in digital holographic microscopy," **Opt. Lett. 2**, 4 (2011).
12. P. Gao, B. Yao, J. Min, R. Guo, B. Ma, J. Zheng, M. Lei, S. Yan, D. Dan, and T. Ye, "Autofocusing of digital holographic microscopy based on off-axis illuminations," Opt. Lett. **37**(17), 3630–3632 (2012).
13. Y. Sun, S. Duthaler, and B. J. Nelson, "Autofocusing in computer microscopy: selecting the optimal focus algorithm," Microsc. Res. Tech. **65**(3), 139–149 (2004).
14. 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).
15. B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," arXiv preprint arXiv:1611.01578 (2016).
16. J. M. Castillo-Secilla, M. Saval-Calvo, L. Medina-Valdès, S. Cuenca-Asensi, A. Martínez-Álvarez, C. Sánchez, and G. Cristóbal, "Autofocus method for automated microscopy using embedded GPUs," Biomed. Opt. Express **8**(3), 1731–1740 (2017).
17. Domain Data Part 1 & 2, and Channel Data for "Multi-domain deep learning for single-frame rapid autofocusing in whole slide imaging," [retrieved 8 March 2018], https://doi.org/10.6084/m9.figshare.5936881.

1. Introduction

High-density solid-state detector technology, coupled with affordable, terabyte-scale data storage, has greatly facilitated the development of whole slide imaging (WSI) instruments. In the biological realm, high-throughput digital imaging has undergone a period of exponential growth catalyzed by changes in imaging hardware and the need for big-data-driven analysis. In the medical realm, there has been an upsurge in worldwide attention on digital pathology [1], which converts tissue sections into digital slides that can be viewed, managed, and analyzed on computer screens. A major milestone was accomplished in 2017 when the US Food and Drug Administration approved Philips' WSI system for the primary diagnostic use in the US [2]. Converting microscope slide into digital images also enable teleconsultations and adoption of artificial intelligence technologies for disease diagnosis. The new generation of pathologists trained on WSI systems and the emergence of artificial intelligence in medical diagnosis promises further growth of this field in the coming decades.

A typical WSI system uses a 0.75 numerical aperture (NA), 20X objective lens to acquire high-resolution images of the sample. The acquired images (tiles) are then aligned and stitched together to produce a complete and seamless image of the entire slide. The depth of field of such a high NA objective lens is less than 1 μm, and thus, it is challenging to acquire in-focus images of different tiles of a sample with uneven topography. Autofocusing issue has been often cited as the culprit for poor image quality in digital pathology [5, 6]. This is not because autofocusing is difficult to do, but rather because of the need to perform accurate autofocusing at high speed and on the fly with the acquisition process.

Conventional reflection based autofocusing methods cannot handle tissue slides with topography variation above the reference glass interface [4]. In current WSI systems, autofocusing solutions include focus map surveying, dual camera setups, optical coherent tomography (OCT) for depth sensing, among others. The focus map surveying approach

Biomedical Optics EXPRESS

creates a focus map prior to scanning. For each point in the map, it typically moves the sample to different focal positions and acquires a z-stack. The best focal position is recovered by maximizing the image contrast of the acquired z-stack. This process is then repeated for other tiles and it is common to skip every 3-5 tiles to save time. Recently, we have demonstrated an implementation with two LEDs for focus map surveying without axial scanning [3]. The dual camera approach employs a secondary camera to acquire images for the autofocusing purpose [4–6]. It requires no focus map surveying and the images can be captured on the fly without axial scanning. However, the use of an additional camera and its alignment to the microscope may not be compatible with most existing WSI platforms. The OCT approach performs depth scan of the sample in high speed. However, it requires expensive and complicated Fourier-domain OCT hardware.

Here we explore the use of deep convolution neural networks (CNNs) to predict the focal position of the acquired image without axial scanning. We compare the autofocusing performance with three illumination settings: 1) incoherent Kohler illumination, 2) partially coherent illumination with two plane waves, and 3) partially coherent illumination with one plane wave. We acquire ~130,000 images with different defocus distances as the training data set. Different defocus distances lead to different spatial features in the captured images. However, solely relying on the spatial information leads to a relatively bad performance of the autofocusing process. It is better to extract defocus features from transform domains of the acquired image. For incoherent illumination, Fourier cutoff frequency is directly related to the defocus distance. Similarly, autocorrelation peaks are directly related to the defocus distance for two-plane-wave illumination. In our implementation, we use the spatial image, the Fourier spectrum, the autocorrelation of the spatial image, and combinations thereof as the inputs for the CNNs. We show that the information from the transform domains can improve the performance and robustness of the autofocusing process. The resulting focusing error is ~0.5 μm, which is within the 0.8-μm depth-of-field range. The reported approach requires little hardware modification for conventional WSI systems and the images can be captured on the fly without focus map surveying. It may find applications in WSI and time-lapse microscopy. The transform- and multi-domain approaches may also provide new insights for developing microscopy-related deep-learning networks. We have made our training and testing data set (~12 GB) open-source for the broad research community.

The contribution of this paper is in threefold. First, we demonstrate the use of deep CNNs for single-frame rapid autofocusing in WSI. Different from the previous implementations, our approach requires neither a secondary camera nor focus map surveying. Second, we employ the transform- and multi-domain approaches to improve the accuracy and robustness of the proposed approach. The use of transform-domain information leads to a better autofocusing performance. To the best of our knowledge, this strategy is new for microscopy applications and may provide new insights for developing microscopy-related deep-learning networks. Third, we have made our ~12 GB training and testing data set open-source for the broad research community. The interested reader can explore better strategies for rapid autofocusing.

This paper is structured as follows: in Section 2, we discuss the deep neural network model we employ in this work. We also discuss the three different illumination conditions under investigation. In Section 3, we compare the performances with spatial-only inputs, transform-domain-only inputs, and multi-domain inputs. We also test the trained CNNs for acquiring whole slide images of different types of samples. Finally, we summarize the results and discuss future directions in Section 4.

2. Methods

The employed deep residual network architecture is shown in Fig. 1. It has been shown that deep residual networks achieve state-of-the-art performance in many image classification and processing applications [7–10]. In Fig. 1, the input to the network is a sample image captured

Biomedical Optics EXPRESS

at a defocus position. This input image first passes through a convolution layer labeled as 'Conv1' in Fig. 1, which contains 64 filters and each filter is of 7 by 7 pixels with a stride of 2 and padding of 3 ('64_7_2_3′ in 'Conv 1'). After transmitting through a maximum pooling layer with a stride of 2, it successively passes through 4 residual blocks [8] labeled as 'Conv2', 'Conv3′, 'Conv4', and 'Conv5′ in Fig. 1. The label ' × 3′ on top of 'Conv 2' block means repeating the block for three times. The signal then passes through a 7 by 7 average pooling layer with a stride of 7 and a fully connected layer. The output of the network is a regression layer and it predicts the defocus distance of the sample.

The training data was acquired using a Nikon Eclipse motorized microscope with a 0.75 NA, 20X objective lens. The samples for training are 35 research-grade human pathology slides with Hematoxylin and eosin stains (Omano OMSK-HP50). The images were acquired using a 5-megapixel color camera with 3.45 μm pixel size (Pointgrey BFS-U3-51S5C-C). We have tested three different illumination conditions for the autofocusing process: 1) regular incoherent Kolner illumination condition with the illumination NA matching to the detection NA, 2) partially coherent illumination with two plane waves (dual-LED), and 3) partially coherent illumination with one plane wave (one-LED). Kolner illumination is employed in most existing WSI systems. Dual-LED illumination has been recently demonstrated for single-frame focus map surveying with an offset distance [3]. For dual-LED illumination, the captured image contains two copies of the sample and the separation of the two copies is directly related to the defocus distance. Single-LED illumination is similar to that of regular holographic imaging settings. Autofocusing for holographic imaging is also an active research topic [11, 12]. In our implementation, we placed two spatially-confined LEDs at the back focal plane of the condenser lens for partially coherent illuminations. As such, we can switch between 3 different illumination conditions without modifying the setup.

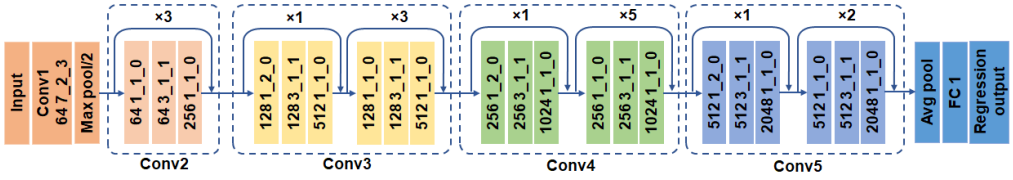


Fig. 1. The architecture of the deep residual network employed in this work. The input for the network is the captured image with an unknown defocus distance. The output of the network is the predicted defocus distance.

In the acquisition process, we acquire a z-stack by moving the sample to 41 different defocus positions in the range from −10 μm to + 10 μm with a 0.5-μm step size. In most cases, the range from −10 μm to + 10 μm is sufficient to cover the possible focus drift of adjacent tiles. This range is also similar to the image-contrast-based methods. We recover the in-focus ground truth by maximizing the Brenner gradient of the z-stack images [13, 14]. For each z-position, we acquire three images with the three illumination conditions discussed above (i.e., three z-stacks for each location of the sample). Figure 2 shows an example of the three z-stacks we captured for the training data set. For the incoherent illumination condition in Fig. 2(a), we can see that the image contrast is higher for the positive defocus direction and this may be due to the asymmetry property of the axial point spread function. For the other two illumination conditions in Fig. 2(b) and 2(c), we take the green channels of the color images to get monochromatic intensity images (the employed LEDs are in green color).

In the training process, we divide the acquired 5-megapixel images into 224 by 224 smaller segments and minimize the difference between the network prediction and the ground-truth defocus position of the training data set. The spatial features of the acquired images are related to the defocus positions of the sample, and this can be seen in Fig. 2. However, solely relying on the spatial features may not be optimal for the autofocusing process. We propose to use or add Fourier spectrum and autocorrelation information as inputs

Biomedical Optics EXPRESS

for the networks. The intuition behind this approach can be explained as follows. For incoherent illumination, the cutoff frequency of the Fourier spectrum is directly related to the defocus distance. For coherent illumination with two LEDs, the Fourier power spectrum contains a fringe pattern whose period is related to the defocus distance, and the image autocorrelation contains two first-order peaks whose locations are related to the defocus distance.

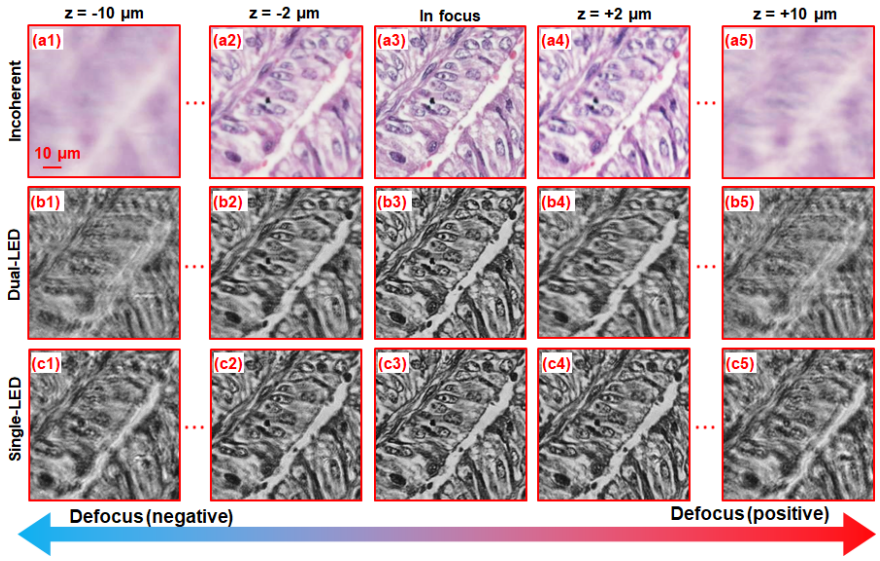


Fig. 2. The three z-stacks for three illumination conditions.

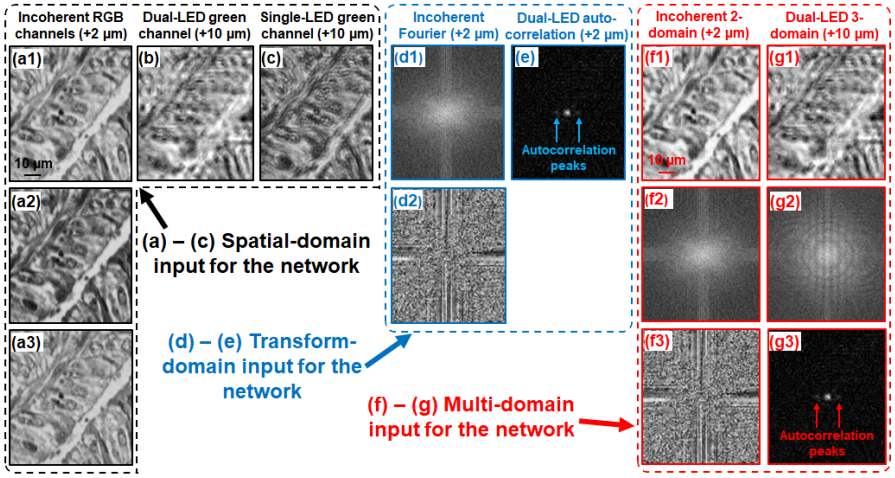


Fig. 3. Comparison between spatial-domain-only input ((a)-(c)), transform-domain-only input ((d)-(e)), and multi-domain input ((f)-(g)) for the networks. (a) The red, green, and blue spatial inputs for the incoherent illumination condition. (b) The single green channel input for the dual-LED illumination condition. (c) The single green channel input for the single-LED illumination condition. (d) The Fourier-domain-only input for the incoherent illumination condition with a Fourier magnitude channel (d1), and Fourier angle channel (d2). (e) The autocorrelation-only input for the dual-LED illumination condition. (f) The two-domain input for the incoherent illumination condition with a spatial intensity channel (f1), a Fourier magnitude channel (f2), and a Fourier angle channel (f3). (g) The three-domain input for the dual-LED illumination condition with a spatial intensity channel (g1), a Fourier magnitude channel (g2), and an autocorrelation channel (g3). All data can be downloaded from Dataset 1 [17].

Biomedical Optics EXPRESS

Figure 3 shows different inputs for the 7 networks. It can be divided into three groups: spatial-domain only inputs (Fig. 3(a)-3(c)), transform-domain-only inputs (Fig. 3(d)-3(e)), and multi-domain inputs (Fig. 3(f)-3(g)). In Fig. 3(a), the input is red, green, and blue spatial channels for the captured incoherent color image. Figure 3(b) shows the single green spatial input for the dual-LED case and Fig. 3(c) shows the single green spatial input for the single-LED case. Figure 3(d) shows the Fourier-domain-only input for the incoherent illumination condition with a Fourier magnitude channel (Fig. 3(d1)) and a Fourier angle channel (Fig. 3(d2)). Figure 3(e) shows the autocorrelation-only input for the dual-LED illumination condition. Figure 3(f) shows the input for the two-domain incoherent illumination case and the channels in Fig. 3(f1)-3(f3) are spatial intensity, Fourier magnitude, and Fourier angle respectively. Figure 3(g) shows the input for the dual-LED illumination case and the channels in Fig. 3(g1)-3(g3) are spatial intensity, Fourier magnitude, and autocorrelation respectively.

In Fig. 3, we did not include the cases of the transform- and multi-domain inputs for the single-LED illumination. The reason is that, the Fourier spectrum and autocorrelation has little correlation with the defocus distance for the single-LED illumination case (the cutoff frequency remains the same for different defocus distances and there is no specific feature in the autocorrelation plot for the defocus distance). As we will discuss later, the deep residual networks with inputs shown in Fig. 3(e)-3(g) give us the best autofocusing performance.

3. Autofocusing performance

With the 7 different inputs shown in Fig. 3, we have trained 7 networks for predicting the defocus distance. The entire training data set contains ~130,000 images (Dataset 1) [17]. The training process is run on a desktop computer with dual Nvidia GTX 1080 Ti graphic cards, an Intel i7-7700k CPU, and 64 GB memory. The networks' weights are learned by using stochastic gradient descent with momentum (SGDM) to minimize the network prediction of the training data set and the ground-truth defocus distance. We empirically set an initial learning rate of 10^{-4} and reduce it 10 times for every 10 epochs. The mini-batch size is set to be 40 images. The training process is terminated when the error for the validation data set starts to increase. The training time ranges from 10 - 30 hours for each of the 7 networks.

To evaluate the performance, we choose two types of samples for testing. The first type of samples is the stained tissue slides from the same vendor (Omano OMSK-HP50) as those used in the training data set (these slides have not been used in the training process). The second type of samples is de-identified H&E skin-tissue slides prepared by an independent clinical lab (the Dermatology Department of the UConn Health Center). In Figs. 4-6, we term the first type of samples as "different samples, same protocol" and the second type of samples as "different samples, different protocol".

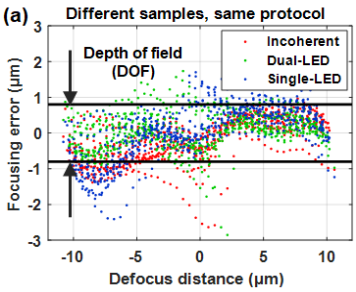
In the testing process, we divide one acquired image into 224 by 224 smaller segments. These segments pass through the trained networks. We then discard 10 outliners from the segment predictions and the remaining predictions (from the small segments) are averaged to give the final defocus distance of the one input image. The reason for discarding outliners is some segments contain mostly empty regions and the predictions from these segments are not reliable. The choice of 10 outliners is based on the assumption that at most 10 segments are empty for each captured image. This assumption is true in most cases we have seen so far.

The strategy of getting rid of outliners is similar to perform teaching evaluation of a course. All students (224 by 224 segments) in the class will give evaluations for the teacher. However, some students (segments with empty regions) are not responsible and always give '0'. Therefore, the final evaluation score is typically based on the median of all evaluation scores (getting rid of outliners) instead of the average. In the left panels of Figs. 4-6, each data point represents the focusing error (y-axis) at a certain ground-truth defocus distance (x-axis).

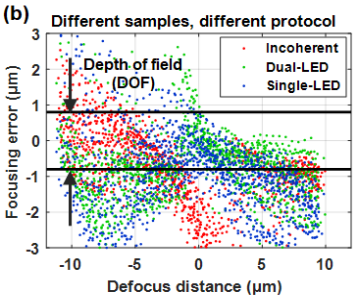
In Fig. 4, we show the autofocusing performance for three networks with spatial-domain only inputs, corresponding to the cases in Fig. 3(a)-(c)). The focusing errors are summarized in the table on the right. There are several observations from Fig. 4. First, the dual-LED

Biomedical Optics EXPRESS

illumination case achieves the best performance for both the type 1 and type 2 samples. The intuition behind this is the separation between the two copies provides direct information for the defocus distance. Second, the performance of type 2 sample is worse than type 1 sample. The reason may be the spatial features of the type 2 samples are new to the networks. It may also justify the need of adding spatially independent features for the networks, such as the Fourier cutoff frequency and autocorrelation peaks. Third, the overall performance of the incoherent network with three color channels is the worst among the three.



Tissue samples	# of tiles	Focusing error (μm)		
		Incoherent RGB	Dual-LED green	Single-LED green
Sample 1	164	0.33 ± 0.25	0.42 ± 0.34	0.65 ± 0.38
Sample 2	82	0.33 ± 0.26	0.49 ± 0.33	0.32 ± 0.26
Sample 3	41	0.37 ± 0.22	0.21 ± 0.20	0.44 ± 0.28
Sample 4	246	0.53 ± 0.28	0.30 ± 0.21	0.57 ± 0.46
Sample 5	82	0.58 ± 0.31	0.46 ± 0.33	0.62 ± 0.39
Sample 6	82	0.87 ± 0.57	0.50 ± 0.48	1.08 ± 0.57
Summary	**697**	**0.50 ± 0.32**	**0.40 ± 0.31**	**0.61 ± 0.39**



Tissue samples	# of tiles	Focusing error (μm)		
		Incoherent RGB	Dual-LED green	Single-LED green
Sample 7	41	1.51 ± 1.02	1.19 ± 1.46	1.80 ± 1.96
Sample 8	123	1.32 ± 1.29	1.57 ± 0.82	2.32 ± 1.51
Sample 9	205	2.69 ± 2.41	1.25 ± 1.96	1.62 ± 2.21
Sample 10	246	2.19 ± 2.15	1.25 ± 0.87	1.87 ± 1.51
Sample 11	246	2.83 ± 3.25	1.40 ± 2.54	1.54 ± 2.03
Sample 12	205	1.00 ± 0.77	1.97 ± 2.16	1.99 ± 2.11
Sample 13	246	2.02 ± 2.48	0.47 ± 0.41	0.90 ± 0.69
Summary	**1312**	**1.94 ± 1.91**	**1.30 ± 1.46**	**1.72 ± 1.72**

Fig. 4. The autofocusing performance for three networks with spatial-domain only inputs. (a) Test on different slides from the same set of samples (slides here have not been used in the training process). (b) Test on different slides prepared by a different clinical lab.

In Fig. 5, we show the autofocusing performance for the two networks with transform-domain-only inputs, corresponding to the cases in Fig. 3(d) and 3(e)). We can see that the dual-LED autocorrelation network has a very good overall performance on the two types of the samples. The focusing error is at least 3 times less than that of the spatial-domain only networks in Fig. 4. In particular, the average focusing errors are within the depth of field of the objective lens.

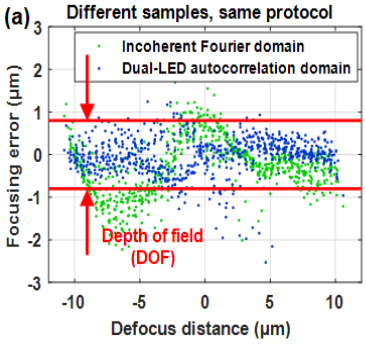
In Fig. 6, we show the autofocusing performance for the two networks with multi-domain inputs, corresponding to the cases in Fig. 3(f) and 3(g)). We can see that the dual-LED three-domain network has a similar performance compared to that of the dual-LED autocorrelation network. The incoherent 2-domain network has the best performance for the incoherent illumination condition.

Based on Figs. 4-6, we can draw the three conclusions: 1) For incoherent illumination condition, the two-domain network has the best performance. 2) For dual-LED illumination condition, the autocorrelation network and the 3-domain network have similar performance. The autocorrelation network performs better on type 2 samples. 3) The networks for dual-LED illumination, in general, perform better than the networks for the incoherent illumination. We also note that, if the defocus value is larger than 10 μm, the networks will predict a relatively large value in the range from −10 μm to 10 μm. The time for getting the predicted focus position from the networks is ~0.04 seconds. For transform-domain and

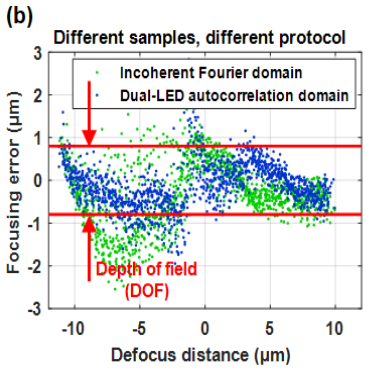
Biomedical Optics EXPRESS

multi-domain networks, another 0.04-0.06 seconds are needed to perform the transform(s). We did not optimize the time in our implementation code.

We have tested the cases of changing illumination NA and changing the objective lens. When we reduce the illumination NA by half, the focusing error using the trained networks increase by 2-3 folds. When we use a new 10X, 0.3 NA objective lens, the network gives a relatively constant prediction. These suggest that if we change the optical configuration, we may need to retrain the network via transferring learning.



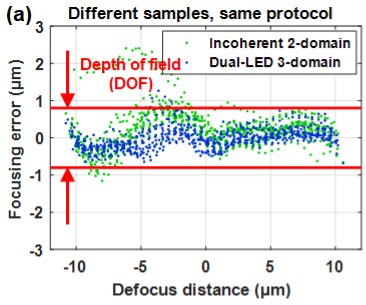
Tissue samples	# of tiles	Focusing error (μm)	
		Incoherent Fourier domain	Dual-LED autocorrelation domain
Sample 1	164	0.61 ± 0.58	0.39 ± 0.43
Sample 2	82	0.70 ± 0.83	0.46 ± 0.37
Sample 3	41	0.53 ± 0.35	0.20 ± 0.20
Sample 4	246	0.50 ± 0.34	0.26 ± 0.19
Sample 5	82	0.63 ± 0.39	0.27 ± 0.25
Sample 6	82	0.70 ± 0.52	0.28 ± 0.23
Summary	697	0.61 ± 0.50	0.31 ± 0.28



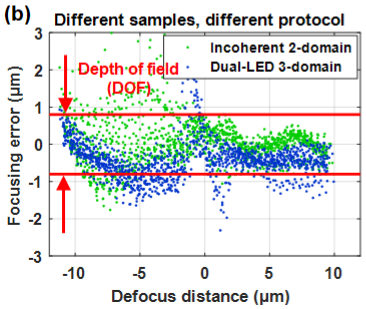
Tissue samples	# of tiles	Focusing error (μm)	
		Incoherent Fourier domain	Dual-LED autocorrelation domain
Sample 7	41	0.94 ± 0.71	0.66 ± 0.58
Sample 8	123	0.99 ± 1.51	0.50 ± 0.45
Sample 9	205	0.63 ± 0.50	0.35 ± 0.27
Sample 10	246	0.77 ± 0.53	0.62 ± 0.50
Sample 11	246	0.72 ± 0.65	0.43 ± 0.36
Sample 12	205	0.52 ± 0.29	0.44 ± 0.27
Sample 13	246	0.62 ± 0.49	0.28 ± 0.22
Summary	1312	0.74 ± 0.67	0.46 ± 0.38

Fig. 5. The autofocusing performance for two networks with transform-domain-only inputs. (a) Test on different slides from the same set of samples (slides here have not been used in the training process). (b) Test on different slides prepared by a different clinical lab.

Biomedical Optics EXPRESS



Tissue samples	# of tiles	Focusing error (μm)	
		Incoherent 2-Domain	Dual-LED 3-Domain
Sample 1	164	0.27 ± 0.18	0.16 ± 0.12
Sample 2	82	0.96 ± 0.86	0.26 ± 0.27
Sample 3	41	0.31 ± 0.22	0.16 ± 0.11
Sample 4	246	0.42 ± 0.24	0.22 ± 0.16
Sample 5	82	0.36 ± 0.29	0.17 ± 0.12
Sample 6	82	0.45 ± 0.24	0.28 ± 0.22
Summary	697	0.46 ± 0.34	0.21 ± 0.17



Tissue samples	# of tiles	Focusing error (μm)	
		Incoherent 2-Domain	Dual-LED 3-Domain
Sample 7	41	0.48 ± 0.32	0.80 ± 0.68
Sample 8	123	1.03 ± 1.50	0.52 ± 0.57
Sample 9	205	0.28 ± 0.28	0.52 ± 0.30
Sample 10	246	0.38 ± 0.38	0.73 ± 0.47
Sample 11	246	0.43 ± 0.69	0.52 ± 0.39
Sample 12	205	0.85 ± 0.73	0.78 ± 0.39
Sample 13	246	0.29 ± 0.22	0.33 ± 0.20
Summary	1312	0.53 ± 0.59	0.59 ± 0.43

Fig. 6. The autofocusing performance for two networks with multi-domain inputs. (a) Test on different slides from the same set of samples (slides here have not been used in the training process). (b) Test on different slides prepared by a different clinical lab.

In Fig. 7, we compare the performance between the spatial-domain only incoherent network and the spatial-Fourier domain incoherent network. Since the spatial features are new to the network (Fig. 7(a)), the spatial-domain network fails to predict the defocus distance in the orange curve in Fig. 7(c). The spatial-Fourier domain network, on the other hand, uses additional Fourier spectrum feature in Fig. 7(b), in which the cutoff frequency is directly related to the defocus distance. The performance of the 2-domain network is shown in the pink curve in Fig. 7(c) and it is more robust for new spatial features it has not seen before.

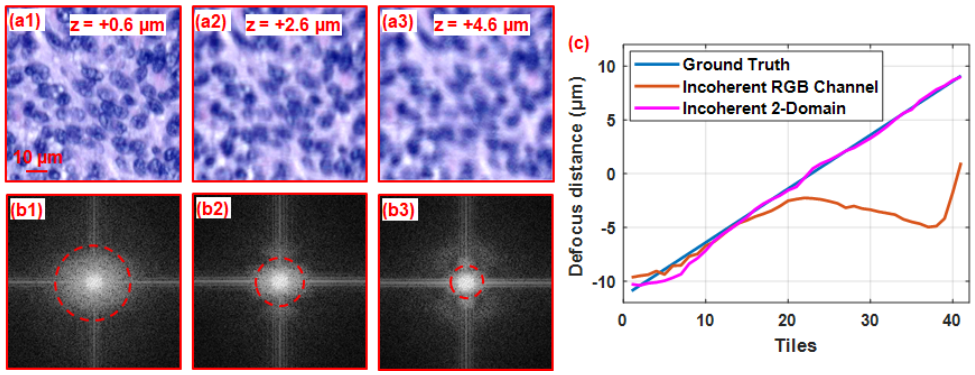


Fig. 7. Comparison between the spatial-domain only incoherent network and two-domain incoherent network. (a) Spatial features at different defocus distances. (b) Fourier-spectrum features at different defocus distance. (c) The predictions of the two networks.

Biomedical Optics EXPRESS

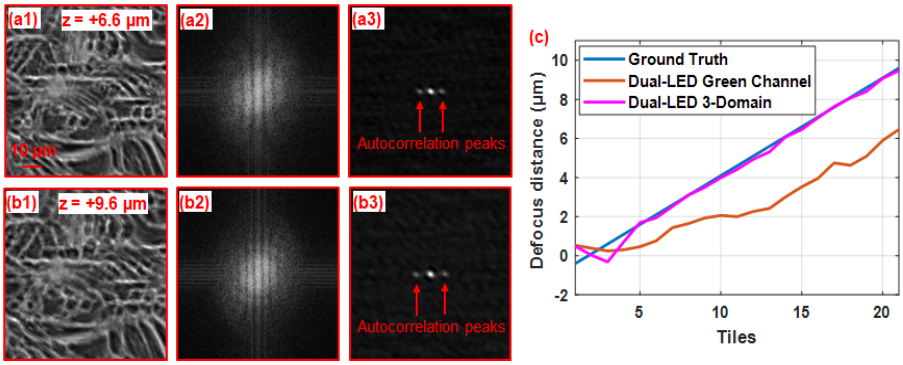


Fig. 8. Comparison between the spatial-domain only dual-LED network and the three-domain dual-LED network. Spatial, Fourier and autocorrelation features at (a) z = 6.6 μm and (b) z = 9.6 μm. (c) The predictions of the two networks.

Likewise, we show an example in Fig. 8 to compare the performance between the spatial-domain only dual-LED network (orange curve in Fig. 8(c)) and the three-domain dual-LED network (pink curve in Fig. 8(c)). For dual-LED illumination, the autocorrelation channel contains two first-order peaks and the distance between these two peaks is directly related to the defocus distance, as shown in Fig. 8(a3) and 8(b3). However, if the defocus distance is too small, the first order peaks cannot be separated from the central peak. The employed three-domain network is able to combine the information from different domains and make the best prediction of the defocus distance, as shown in the pink curve in Fig. 8(c).

In Fig. 9, we tested the use of the two-domain incoherent network to perform whole slide imaging. Figure 9(a) shows the whole-slide image of a type 1 sample and the focus error map is shown in Fig. 9(c1). Figure 9(b) shows the whole-slide image of a type 2 sample and the focus error map is shown in Fig. 9(c2). For both cases, 99% of the focus errors are less than the depth of field of the employed objective lens. The proposed networks may provide a new solution for WSI with neither focus map surveying nor a secondary camera.

Biomedical Optics EXPRESS

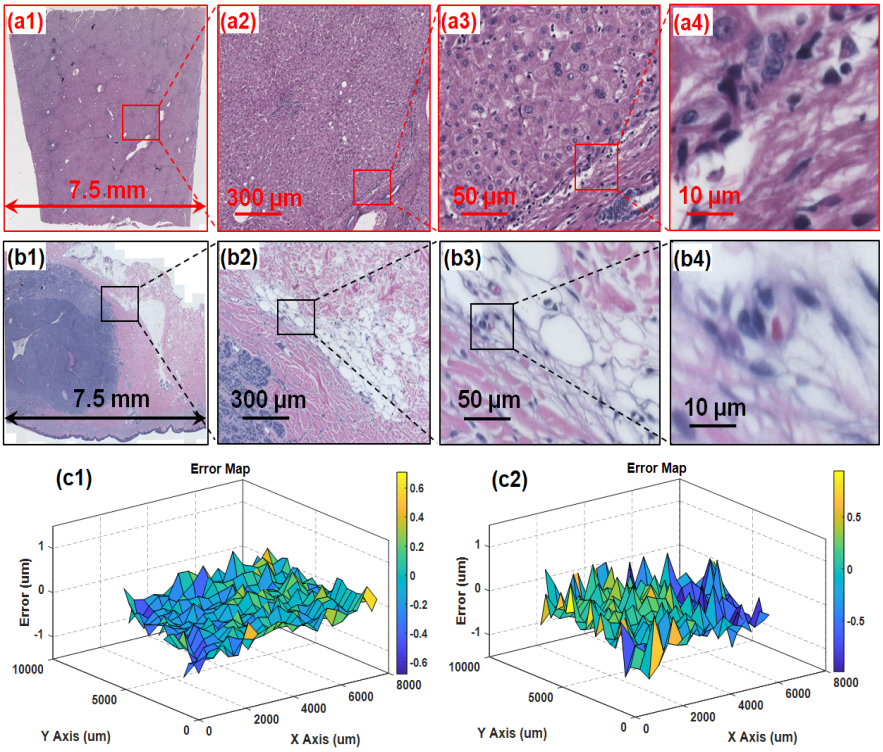


Fig. 9. Test of the two-domain incoherent network for whole slide imaging. (a) The captured whole-slide images of a type 1 sample (a) and type 2 sample (b). (c1) The focus error map for (a). (c2) The focus error map for (b).

4. Discussion

In summary, we report the use of deep residual networks to predict the focus position of the acquired image. Different from conventional CNN implementation which relies on the spatial features of the input images, we explore the use of Fourier spectrum and image autocorrelation as the input channels for the networks. We discuss and compare the performance with three different illumination conditions. For incoherent illumination condition, the two-domain network has the best performance. For dual-LED illumination condition, the autocorrelation network and the 3-domain network have similar performance. For the best networks, the average focusing error is about two times smaller than the depth of field of the employed objective lens. Different from the previous autofocusing approaches, the reported approach requires little hardware modification for existing WSI systems and the images can be captured on the fly with neither a secondary camera nor focus map surveying. The strategy of using transform- and multi-domain information for microscopy imaging, to the best of our knowledge, is new and may provide new insights for developing microscopy-related deep-learning networks.

Some of the findings in our work are counterintuitive. For example, one may think that even we know the sample is defocused by 1 μm, it is difficult to tell it is in the positive or negative direction. This difficulty leads to the use of a sample offset distance in the previous implementation [3], and as such, a focus map surveying process is needed. In this work, we show that the deep learning network is able to recognize the subtle spatial-feature difference under different defocus directions in Fig. 2(a) (due to the asymmetric axial point spread function of the objective lens).

Biomedical Optics EXPRESS

The reported approach may also find applications in focus drift correction in time-lapse experiments. The existing solution is based on laser reflection method which requires the user to choose an offset distance to a reference surface (for dry objectives, the reference surface is the air-dish interface). The offset distance may vary for different locations because the thickness of the dish is not uniform. With proper training, the reported dual-LED networks may be able to automatically pick the best focus position based on the transform- or multi-domain information input. This may be useful for long-term time-lapse cell culture imaging since one can generate coherent contrast of transparent samples using oblique illumination from the two LEDs. The wavelength of the LED can be chosen based on the passband of the emission filter.

We also note that, for some specific applications, the samples have very similar spatial features across the entire slide (blood smear and Pap smear samples). In this case, we can capture a small amount of training data and perform transfer learning of the reported networks.

We envision several future directions of our work. First, other network architectures can be used for better autofocusing performance. Dilated convolution can be used to expand the receptive field. An optimal neural network architecture can also be designed by the reinforcement learning approach [15]. Second, a better strategy can be used in predicting the focus position of the captured image. In the current implementation, we predict the focus position based on the captured image. One improvement is to use the previous focus positions of other segments to better predict current focus position. Another neural network can be used for this purpose. The input of this new neural network is the previous and current predictions from the reported networks in this work. The output of this new neural network is a new prediction of the focus position of the current segment based on all information around this segment. Third, the reported approach can be implemented on an embedded GPU integrated system [16]. Fourth, the gap between the same protocol and the different protocol samples stems from the domain adaptation problem in deep learning. How to minimize this gap is an important future direction.

Appendix

We provide the training and testing dataset for the 7 networks: Dataset 1 (~130,000 images in total and ~12 GB in size) [17]. The name of the folder provides the information of the illumination condition and the input channels. For example, 'train_dualLED_3domains' means it is training data for dual-LED illumination condition with 3-domain inputs. The name of the image file provides the information of the ground-truth defocus distance. For example, 'Seg1_defocus-650.jpg' means it is from segment 1 and the ground-truth defocus distance is −650 nm.

Funding

National Science Foundation (1510077, 1555986, 1700941); NIH (R21EB022378, R03EB022144).

Disclosures

G. Zheng has the conflict of interest with Clearbridge Biophotonics and Instant Imaging Tech, which did not support this work.