Research Article | Vol. 12, No. 9 / 1 Sep 2021 / Biomedical Optics Express 5644
Biomedical Optics EXPRESS

Low-cost whole slide imaging system with single-shot autofocusing based on color-multiplexed illumination and deep learning

KAIFA XIN,[1,2] SHAOWEI JIANG,[3] XU CHEN,[1,2] YONGHONG HE,[1] JIAN ZHANG,[4] HONGPENG WANG,[5] HONGHAI LIU,[5] QIN PENG,[6] YONGBING ZHANG,[5,7] AND XIANGYANG JI[2,8]

[1]*Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, 518055, China*
[2]*Department of Automation, Tsinghua University, Beijing, 100084, China*
[3]*Department of Biomedical Engineering, University of Connecticut, Storrs 06269, USA*
[4]*Shenzhen Graduate School, Peking University, Shenzhen, 518055, China*
[5]*Harbin Institute of Technology (Shenzhen), Shenzhen, 518055, China*
[6]*Institute of Systems and Physical Biology, Shenzhen Bay Laboratory, Shenzhen, 518132, China*
[7]*ybzhang08@hit.edu.cn*
[8]*xyji@tsinghua.edu.cn*

Abstract: Recent research on whole slide imaging (WSI) has greatly promoted the development of digital pathology. However, accurate autofocusing is still the main challenge for WSI acquisition and automated digital microscope. To address this problem, this paper describes a low cost WSI system and proposes a fast, robust autofocusing method based on deep learning. We use a programmable LED array for sample illumination. Before the brightfield image acquisition, we turn on a red and a green LED, and capture a color-multiplexed image, which is fed into a neural network for defocus distance estimation. After the focus tracking process, we employ a low-cost DIY adaptor to digitally adjust the photographic lens instead of the mechanical stage to perform axial position adjustment, and acquire the in-focus image under brightfield illumination. To ensure the calculation speed and image quality, we build a network model based on a 'light weight' backbone network architecture-MobileNetV3. Since the color-multiplexed coherent illuminated images contain abundant information about the defocus orientation, the proposed method enables high performance of autofocusing. Experimental results show that the proposed method can accurately predict the defocus distance of various types of samples and has good generalization ability for new types of samples. In the case of using GPU, the processing time for autofocusing is less than 0.1 second for each field of view, indicating that our method can further speed up the acquisition of whole slide images.

1. Introduction

Whole slide imaging (WSI) aims to get the digital representation of pathology slides containing tissue specimen, which are originally observed by a pathologist through a microscope, time consumingly and workload intensively. The concept of WSI was first developed based on a robotic microscope in late 1990s [1] by Ferreira and Joel Saltz, who designed a software system to indicate the research and development ideas of WSI. Due to the rapid expansion of artificial intelligence in recent years, digital pathology has ushered in a period of rapid development. Simultaneously, research on WSI has received more attention. A landmark is that U.S. Food and Drug Administration approved the whole slide imaging system designed by Philips for digital diagnostic aids in 2017.

In order to acquire the high spatial resolution digital images, a high numerical aperture (NA) objective lens is usually employed in WSI systems. Under the same magnification of objective

#428655 https://doi.org/10.1364/BOE.428655
Journal © 2021 Received 28 Apr 2021; revised 30 Jul 2021; accepted 3 Aug 2021; published 16 Aug 2021

Biomedical Optics EXPRESS

lens, larger NA means smaller depth of field (DOF). However, the small DOF of objective lens poses a great challenge to focus accurately in a new field of view (FoV) during the scanning process. Thus, autofocusing is a critical step in WSI acquisition. In previous WSI settings, autofocusing methods can be mainly categorized into three groups: 1) pre-scan focus map approach, 2) real-time reflective autofocusing, and 3) real-time image-based autofocusing [2].

In traditional WSI settings, a z-stack, corresponding to the tissue images at different focal planes is acquired for each FoV, which is time consuming and storage intensive [1]. A figure of merit is then applied to the images of stack to figure out which image is acquired in the best focal plane. Focus map with Delaunay triangulation algorithm, which determines the focal plane on FoVs discretely distributed and interpolates the rest, is applied to reduce the time consumption in acquiring z stacks by reducing the number of acquisition points [3]. Yazdanfar *et al.* proposed an empirical fitting model that applies a Lorentzian function for Brenner gradient focus measure to minimize the number of images having to acquire within a stack. Applying this model, only 3 images are needed to calculate the best focal plane [4].

Some researches focus on the implementation of reflective-based autofocusing. In 2006, Y.Liron illustrated a laser based autofocusing method with confocal pinhole setup [5]. A two-stage search algorithm is introduced to determine the precise focal plane. Although this method can perform precise autofocusing, axial scanning is still needed to calculate the trace curve. G. Reinheimer *et al.* proposed triangulation concept for microscopy illumination, which projects the illumination light to the sample with an incident angle and measures the lateral displacement of the reflected beam [6].

Recently, real-time autofocusing has been used in WSI platforms by adding additional components. R. R. Mckay and M. C. Montalto proposed independent dual sensor scanning system for real-time autofocusing in 2011 [3]. During image acquisition, while the imaging camera reading out the high-resolution image of the previous FoV, the sample moves as well as the autofocusing sensor acquires three images at different focal planes to calculate the focus position of the next FoV. Dong *et al.* proposed to acquire the best focal plane through a tilted sensor in 2005 [7]. Imaging on tilted focusing sensor is non-uniform defocused and the defocus distance can be inferred from the pixels between the parfocal point and the highest-contrast point. A. Kinba *et al.* described the phase detection autofocusing method, which divides the incoming light into two parts and calculates the pixel shift to infer the defocus distance [8]. Zheng *et al.* proposed real-time autofocusing approaches based on phase detection [9,10] and dual-LED illumination [11–15], which allows continuous sample motion during autofocusing image acquisition. In addition, an OpenWSI system proposed in [15] uses only one sensor for both focusing and image acquisition, which reduces the hardware costs.

Recently, with the rapid development of artificial intelligence, quantities researches applied deep learning for focal plane detection. Jiang *et al.* applied Resnet based convolutional neural network to predict the defocus distance through the transform and multi-domain inputs [16]. Experimental results of [16] show that using brightfield images as inputs of network brings large focusing errors. Especially when facing with pathological slides from different vendor, the mean focusing error can be 2.4 times of DOF, indicating that using brightfield images as input to infer defocus distance has poor generalization ability. Inspired by this, in this paper, we use partial coherent illumination images as input. In this case, both the defocus direction and defocus distance are encoded in the relative position of red and green channels, which contributes to higher robustness and generalization ability. In addition, Dastidar *et al.* employed the difference of two defocused images as the input of a CNN network, defocus distance as the output [17]. However, the preprocessing of the previous two approaches may introduce additional time to the defocus distance prediction. Pinkard *et al.* proposed a method to set an additional off-axial LED as the illumination source, and feed the acquired images to a fully connected Fourier neural network to predict the defocus distance [18]. This method has a larger working distance and a

Biomedical Optics EXPRESS

faster prediction speed, but the focusing error is relative higher when dealing with new sample types. Another application of deep learning in WSI field is to virtually refocus blurry out-of-focus images into in-focus images. Wu *et al.* trained a deep neural network to virtually refocus a two-dimensional fluorescence image onto user-defined three-dimensional(3D) surfaces within the sample [19]. Y. Luo and A. Ozcan proposed the deep-learning based offline autofocusing network, termed Deep R to rapidly and blindly autofocus a microscopy image [20].

This paper described a low cost WSI platform and proposed a robust single-shot autofocusing method, which utilizes the deep learning to estimate the defocus distance. We build the model based on a 'light weight' backbone network architecture-MobileNetV3. The key contributions of this paper are:

- We reported a low-cost WSI scheme based on a photographic lens instead of the tube lens and a programmable LED array for sample illumination. We employ a customized Canon EF mount to industrial camera C mount adapter ring to provide a low-cost solution for subsequent research on photographic lens focusing adjustment.
- We propose the use of deep learning to estimate the defocus distance based on partially coherent illumination images with accurate single-shot autofocusing. In addition, the proposed 'light weight' model can maintain fast computing speed even when used in embedded devices.
- The experimental results show that the proposed method has good robustness for under-stained and thick samples and generalization for new sample types, and the processing time for focusing position estimation is less than 0.1s.

This paper is organized as follows: in section 2, we introduce the employed low-cost WSI system and the deep learning method for autofocusing. In section 3, we present the experimental results on various kinds of samples. Finally, we conclude the paper in section 4.

2. Low-cost WSI based on color-multiplexed illumination and deep learning

2.1. Hardware platform

Figure 1(a) shows the low-cost WSI platform, which is composed of three parts: the imaging system, the illumination system, and the electronically controlled translation system. In the imaging system, we use a 20-megapixel color camera (ImageSource DFK 33UX183, 2.4 μm pixel size) to acquire the digital images. An Olympus objective lens (20X, 0.5NA) and a photographic lens (Canon EF 100mm f2.8L Macro IS USM) are used in the proposed microscopy system. We use a programmable LED array (Adafruit DotStar High Density 8×8 Grid) as the illumination source. Each LED is individually addressable and drivable based on the embedded microcontroller which allows us to switch between the incoherent illumination and partial coherent illumination. On the basis of conventional microscopy system, we only need to replace the illumination equipment, which is composed of LED array, customized diffuser and heat dissipation module, to achieve focusing tracking with the proposed method. It is convenient to apply our method on conventional microscopy system.

Figure 1(b1) shows the schematic diagram of the system when a red and a green LEDs are turned on and illuminate the sample from two opposite incident angles. Under this illumination mode, the acquired color image is shown in Fig. 1(b2). If the sample is placed at a defocus position, there will be an interval between the red channel and green channel of the image, which contains the information of defocus distance. For a certain defocus distance, a larger illumination angle between the red and green LED indicates a larger displacement between the red and green channel of Fig. 1(b2), while a smaller illumination angle leads to a smaller translational shift [21]. Based on the translational shift, defocus distance can be figured out. After the focus tracking

Biomedical Optics EXPRESS

2.1 Hardware platform

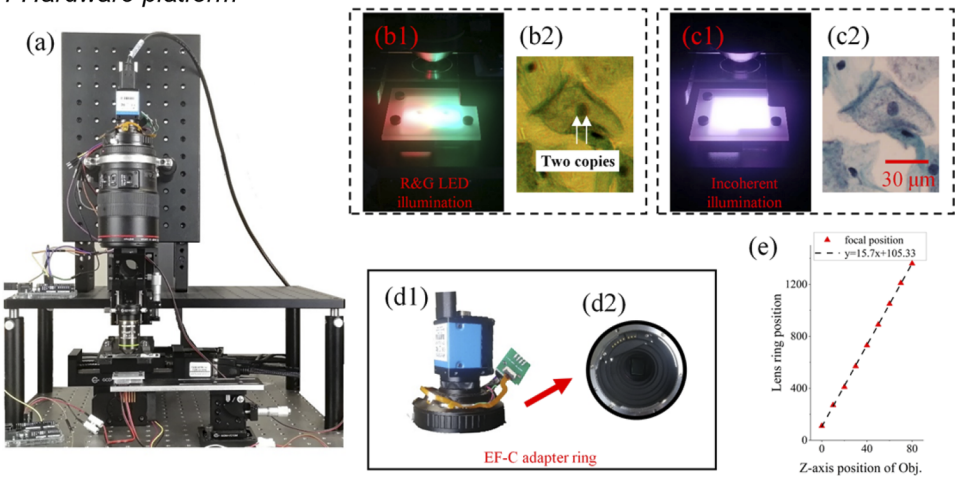


Fig. 1. (a) Low-cost whole slide imaging platform. (b) Schematic diagram of the platform with color-multiplexed illumination by turning on a pair of red and green LEDs. (c) Schematic diagram of the system with incoherent illumination by turning on all LEDs. (d) Customized lens adapter compatible for EF-mount Canon lens to C-mount industrial camera (e) The measured calibration curve between lens ring position and axial defocus distance of the sample.

process, we switch the LED array to incoherent illumination mode by turning on all LEDs as shown in Fig. 1(c1) and acquire a brightfield image as shown in Fig. 1(c2). For the initial state of the acquisition process, we turn on two green LEDs to illuminate the sample from opposite incident angles and capture two images, respectively. Then the pixel shift between these two images, which is used to estimate the focal plane of the first FoV, can be calculated by locating the maximum point of the cross-correlation plot.

We customized an adapter ring, as is shown in Fig. 1(d) for EF-mount photographic lens to C-mount industrial camera. The main task of making the adapter ring is to install pogopins on appropriate positions of a commercially available EF-C adapter, which is made of aluminum alloy and has no contacts integrated inside, to lead out the control wirings. We install a Canon 550D camera-body-side contact module to the adapter ring to achieve this, which is a simpler solution. The OpenWSI system controls the photographic lens by opening the lens and leads out the control wiring from the inside [15]. This intrusive method will cause damage to the lens and may affect the use of the lens in other experiments. The adapter ring we designed can avoid this problem elegantly, and allows us to perform precise axial positioning control at a low cost. We use the Serial Peripheral Interface (SPI) of the Arduino microcontroller to simulate the Canon camera body sending control commands. Thus, the ultrasonic motor ring inside the lens can be controlled to turn to different positions. The measured calibration curve between the lens ring position and the axial defocus distance of the sample is shown in Fig. 1(e). The slope of the curve shows that in our setting, objective lens moving along z axis for 1 μm corresponds to ultrasonic motor adjusting 15.7 lens ring position.

The main principles of the illumination system are shown in Fig. 2. A diffuser is placed between the LED array and the sample to provide incoherent illumination. Two holes are required at the diffuser plane for partial coherent illumination during color-multiplexed image acquisition. The schematic diagram of the illumination system is shown in Fig. 2(a). According to the

Biomedical Optics EXPRESS

geometric relations, the position of the hole $x1$ is given as:

$$x1 = \frac{d1 \cdot x2}{d1 + d2} \tag{1}$$

where $d1$ is the distance between the sample and the diffuser, $d2$ is the distance between the diffuser and the LED array, and $x2$ is half the distance between the red and green LEDs. $x2$, $d1$ and $d2$ can be measured directly and the positions of the holes on the diffuser can also be determined. Under partial coherent illumination, there is a linear relationship between the defocus distance and the pixel shift, as shown in Fig. 2(b) and Eq. (2),

$$p = \frac{D \cdot z}{H \cdot psize} \tag{2}$$

where D is the distance between the red and green LEDs. H is the distance between the LED array and the pathological slide. z represents the defocus distance. p represents the pixel shift and $psize$ represents the pixel size.

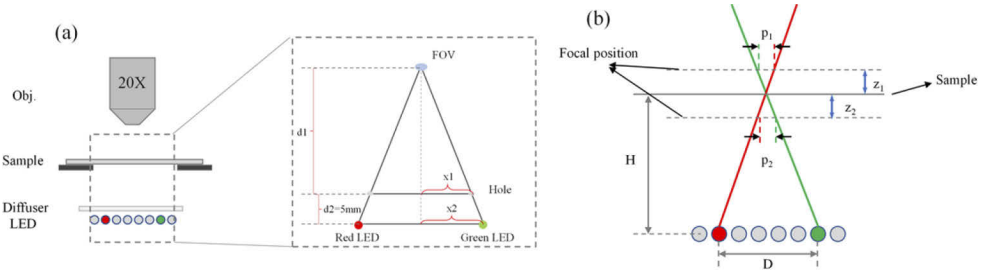


Fig. 2. (a) Schematic diagram of the illumination system. (b) Relationship between defocus distance and pixel shift.

Under color-multiplexed illumination, the relative position between the red channel and the green channel relates to the defocus direction, as shown in Fig. 3. The red channel is on the left of the green channel when defocusing in the negative direction and the red channel is on the right of the green channel when defocusing in the positive direction. The accuracy of the defocus direction judgement is one of the advantages of our system, which will be described in section 3. There is a clear and direct relationship between the translational shift and defocus distance. As the defocus distance increases, the translational shift between the red and green channels also increases. Compared with extracting defocus distance through brightfield images, there is a more direct relationship between translational shift and the defocus distance, which is more conductive to figure out the defocus distance accurately.

Real time autofocusing can be realized in the case of using the programmable LED array and drivable photographic lens. When the sample moves to a new FoV, the red and green LEDs are turned on to provide color-multiplexed illumination and the captured image is used to figure out the defocus distance. Then the LED array switches to provide incoherent illumination and in-focus brightfield image is acquired after the photographic lens adjusting focal plane to the target position.

2.2. Calibration

We employ a low-cost manual stage for coarse axial adjustment when changing different samples, and perform precise autofocusing via adjustment of the ultrasonic motor ring inside the photographic lens. The calibration curve between the lens ring position and the z-axis position of the objective lens, as shown in Fig. 1(e), is measured by manually adjusting the objective lens

Biomedical Optics EXPRESS

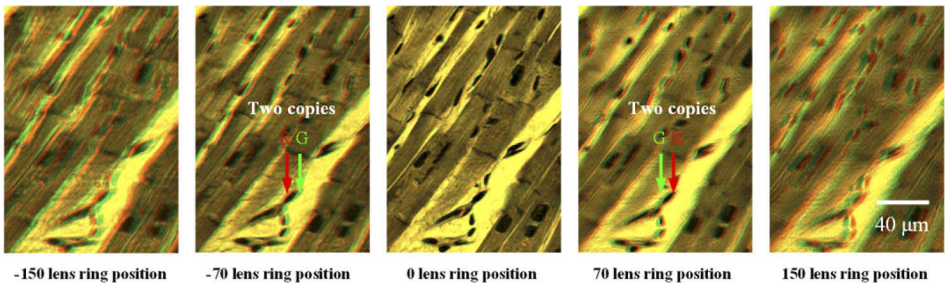


Fig. 3. Relative relationship between the red and the green channel under color-multiplexed illumination at different defocus distance.

to different positions, for which we find the adjustment of the lens ring position. The positioning accuracy of the manual stage reaches 10 μm, which can meet the experimental requirements. We can see that the lens ring position adjustments are in good agreement in a linear relationship with the z-axis positions of objective lens. The slope of the curve manifests that in our setting, objective lens moving along z axis for 1 μm equals to ultrasonic motor adjusting 15.7 lens ring position. We achieve precise focal plane adjustment at a low cost by digitally adjusting the motor ring instead of using an expensive mechanical stage. A two-dimensional electronically driven stage is employed to enable lateral scanning of the pathological slides. Slide holder is mounted on the stage so that we can move the pathological slides to align the objective lens to different FoVs.

Figure 4(a) shows the Brenner gradient of the image stacks acquired by adjusting lens ring position when objective lens is set at different z-axis positions. Brenner gradient is a measurement of average change in gray level between pairs of point separated by two pixels, which was proposed in 1976 by J. F. Brenner [22]. The Brenner gradient is acutely sensitive to focus, monotonically decreasing and symmetric about the peak. The Brenner gradient of a gray scale image is shown in Eq. (3)

$$Brenner(s) = \sum_{m=1}^{M-2} \sum_{n=1}^{N} (s(m,n) - s(m+2,n))^2 \tag{3}$$

where m and n are pixel indexes of images. We first covert the acquired color images into grayscale images, and then calculate the Brenner gradient by Eq. (3).

The extreme point of each curve represents the lens ring position which is best in focus. We mount the objective lens on a low-cost manual stage, so the objective lens can move along z-axis. In Fig. 4(a), for each curve, the objective lens has a displacement of 10 μm. Brightfield images on different focal planes are shown in Fig. 4(b). Image taken closer to the best focal plane have a larger Brenner gradient. Using the relative position of the objective lens on z axis as the horizontal axis and the lens ring position of the extreme point as the vertical axis, the image shown in Fig. 1(e) can be obtained. Through the relationship between the z-axis position of the objective lens and the lens ring position, the proportional relationship between photographic lens adjustment and z-axis adjustment can be determined.

2.3. Deep learning based autofocusing method

To ensure the calculation speed and image quality, we proposed a deep learning based autofocusing method. We build our network based on MobileNetV3 [23] architecture, which is the new generation of MobileNets published by Google AI. This "light weight" architecture enables high inference efficiency and low computing resources, which maintains fast speed on terminal devices. MobileNetV3 introduces light-weight attention modules based on squeeze and excitation into the

Biomedical Optics EXPRESS

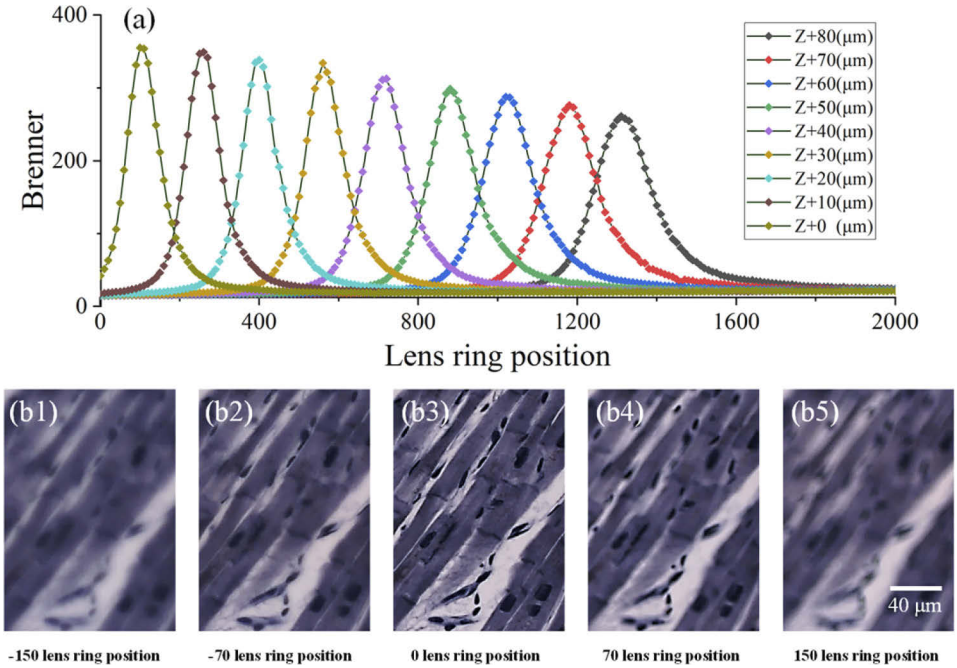


Fig. 4. (a) Brenner gradient of image stacks acquired by adjusting lens ring position when the objective lens is placed at different z-axis positions. For adjacent curves, interval of the objective lens in the z-axis equals 10 μm. (b) Brightfield images of different focal planes obtained by adjusting the ultrasonic motor ring inside the photographic lens.

bottleneck structure compared with MobileNetV2 [24]. Layers also are upgraded with modified swish nonlinearities.

The network structure used in this paper is shown in Fig. 5. The network is composed of three kinds of blocks, namely Squeeze and Excitation (SE) block, Inverted Residual block (IRB) and the block composed of the above two structures (SEIRB). The 'DW Conv' in Fig. 5 means Depth-wise convolution, which use each filter channel only at one input channel. Combined with the following '1 × 1 Conv', Depth-wise separable Convolution uses fewer parameters to achieve the same effect of normal convolution and to reduce the computational cost. The 'SE Module' means 'Squeeze-and-Excitation' architectural unit. The 'SE' module improves the representational power of a network by enabling it to perform dynamic channel-wise recalibration [25]. In 'SE Module', each channel is first 'squeezed' into a numeric value and then reduced by a ratio through a dense layer. Afterwards, weights of each channel are given by another dense layer, and input channels of 'SE Module' are weighted by the weights finally, which is called 'Excitation'.

In our application scenario, the defocus distance is related to the translational shift between channels. SE module applied in the network is able to extract the useful information between channels, which helps to improve the accuracy. The small parameter size of the network enables fast computational speed and availability in embedded devices. The proposed network also provides the best performance among the networks we tested.

For image data preparation, both incoherent illumination focal stacks and partial coherent illumination focal stacks are collected. We capture image stacks within a defocus range of ±160 lens ring position, which is equivalent to ±10 μm. In data acquisition process, we collect z-stacks by commanding ultrasonic ring to 33 different defocus positions in the range between -160 lens

Biomedical Optics EXPRESS

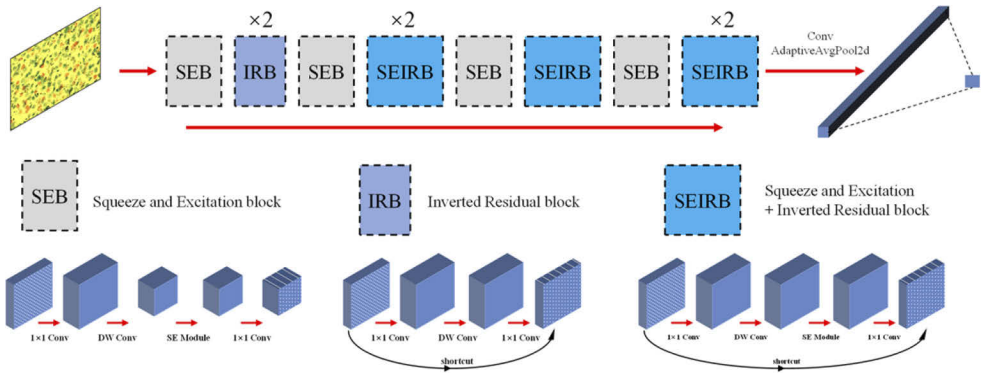


Fig. 5. Network structure. 'SE Module': Squeeze and Excitation. 'DW Conv': Depth-wise convolution. The input for the network is the red and green channels of color-multiplexed image. The output of the network is the defocus distance.

ring position to +160 lens ring position with a 10 lens ring position step size, approximately 0.637 μm step size. We acquire a brightfield image and a partial coherent illumination image at every focal position. In this way, an incoherent illumination stack, as shown in Fig. 6(a), and a color-multiplexed illumination stack, as shown in Fig. 6(c), are obtained at each FoV. The images of these two stacks are one-to-one correspondent.

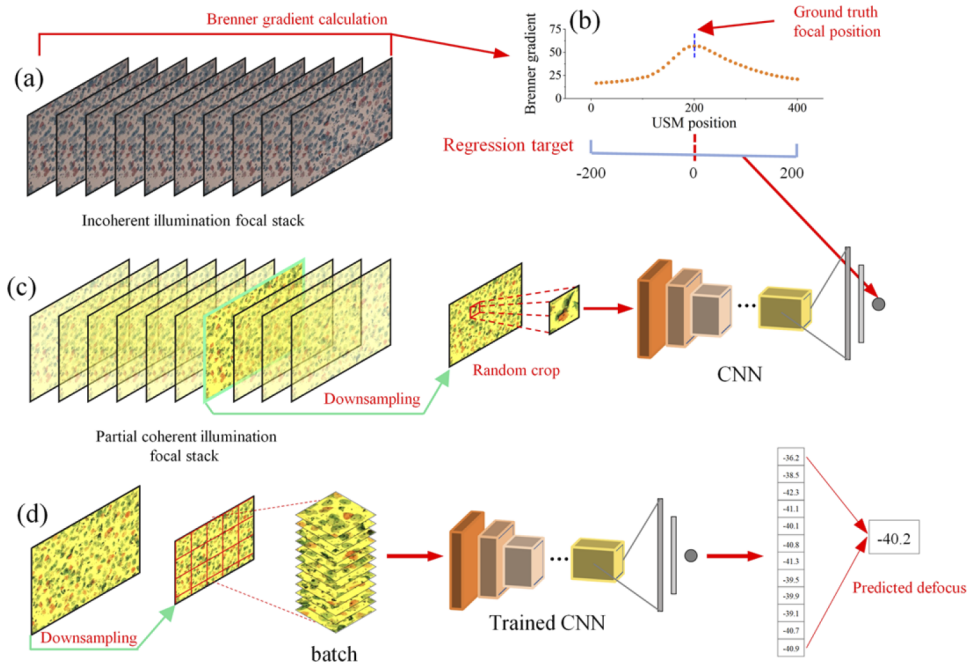


Fig. 6. Data acquisition, training and defocus prediction. (a) Incoherent illumination stack. (b) For each image in the incoherent illumination focal stack, Brenner gradient is calculated to figure out the ground truth focal position (c) In the training process, each image in the color-multiplexed illumination focal stack is down-sampled and cropped as the input of the neural network. (d) In the inference process, each color-multiplexed image is employed to figure out the defocus distance.

Biomedical Optics EXPRESS

The incoherent illumination stacks are used to determine the ground truth focal position and the defocus distance of each image. The color-multiplexed illumination stacks are fed into the convolutional neural network to predict the defocus distance. For each image in incoherent focal stack, we calculate the Brenner gradient and consider the position where the brightfield image having the largest Brenner gradient as the ground truth focal position, as shown in Fig. 6(b) and Eq. (4) [22].

$$Index_{Bre\_max} = arg \max_{0<i<N}(Brenner(B_i)) \tag{4}$$

where N is the number of images in a stack, B_i represents the i^{th} image of the brightfield image stack. Then the defocus distance of each image in incoherent illumination stack can be obtained according to the step size and the index difference from in-focus position:

$$Defocus(B_i) = Szpl \cdot (i - Index_{Bre\_max}), \ \ (1<i<N) \tag{5}$$

where $Szpl$ represents the step size of photographic lens adjustment during image acquisition. In this paper, we adjust the focal position with 0.673 μm step size with the 1.34 μm DOF. Consequently, regarding the position of image with largest Brenner gradient as ground truth focal position will not cause substantially large deviation. In the case of a large training set, this deviation will be neutralized. After the ground truth focal position is acquired, the defocus distance of each image in the incoherent illumination stack is obtained. Then, the defocus distance of each image in the partial coherent illumination stacks is also known because of the one-to-one correspondence.

In training process shown in Fig. 6(c), we crop each color-multiplexed illumination image into 512×512 tiles, and use the red and green channels as the input of the network. We remove the blue channel to avoid the redundant information of color crosstalk, which may affect the accuracy of prediction. During inference process shown in Fig. 6(d), we split the acquired color-multiplexed image into 12 non-overlapping 512×512 sub-images. The 12 sub-images are put into the trained network as a batch to get 12 outputs. We remove the largest two and the smallest two of the 12 outputs and take the average as the final defocus distance.

The machine used for training the network is equipped with an Intel Xeon E5-2650 processor, a RTX 2080Ti graphic card and runs on ubuntu 16.06. We set the learning rate to 0.001 in the first 40 epochs, and then the learning rate drops to 90 percent after each epoch in the next 30 epochs. The model after each epoch will be used to test the accuracy on the validation set and we will update the saved model parameters when the accuracy improves. If there is no improvement on accuracy on validation set for 4 epochs, training will be terminated early. We take *MSELoss* as the loss function, as shown in Eq. (6).

$$MSELoss(y_i, \bar{y}_i) = (y_i - \bar{y}_i)^2 \tag{6}$$

where i refers to the i^{th} index of the input images, $\bar{y}_i$ refers to the output, y_i refers to the actual defocus distance.

3. Experimental results

3.1. Dataset

We obtained 112 pathological slides of different types, including hematoxylin-eosin (HE) stained cell slides, HE stained tissue slides, HE stained mouse myocardial tissue slides, HE stained human bacterial myocarditis tissue slides, and hematoxylin-stained human myocarditis tissue slides.

We capture stacks by the proposed hardware system with a 20X Olympus objective lens and a 20-megapixel camera. For each FoV, 33 images under color-multiplexed illumination and 33 images under incoherent illumination are captured with defocus distance varying from -160 to

Biomedical Optics EXPRESS

160 lens ring position with a step size of 10 lens ring position, equivalent to ~0.637 μm. For real-time focusing strategy, the in-focus position between two adjacent tiles is no more than 10 μm, so the -160 to 160 lens ring position range is sufficient for the reported system. A total number of 23,171 20-megapixel images are available for training. The data is split into a 19, 695 training set and a 3, 476 validation set. Testing set consists of 6158 20-megapixel images, which can be divided into two categories based on whether the sample type appearing in the training set.

3.2. Performance on test set

We validate our method by comparing with mutual information (MI) maximization algorithm. MI has been widely used for nonrigid multimodality image registration. MI measures how much information one random variable contains about the other random variable. A significant advantage of this algorithm is its capability of dealing with images that is not quite analogical [21]. If X and Y are the random variables which represent the intensities of two images, the MI $I(X, Y)$ between two images is given by:

$$I(X, Y) = H(X) + H(Y) - H(X, Y) \tag{7}$$

where $I(X, Y)$ is the mutual information of X and Y, $H(X)$ and $H(Y)$ represent the entropy of X and Y respectively, $H(X, Y)$ is the joint entropy of X and Y. S. Jiang proposed to calculate the displacement between the red channel and green channel with subpixel resolution by maximizing mutual information [21]. The displacement is proportional to the defocus distance of the sample.

The samples in the test set can be divided into three categories: cell samples, human tissue samples and mouse myocardial samples. Figure 7(a) shows the pixel shift calculated by MI maximization algorithm. For most samples, pixel shift being near the fitted curve shows that MI maximization algorithm can be used to calculate the pixel shift. Then, the defocus distance can be figured out because of the linear relationship between defocus distance and pixel shift. Our method is applied to the test set and the corresponding focusing error is shown in Fig. 7(b). The mean focusing error is approximately 0.31 μm, which is well within the ±1.34 μm DOF, indicating that our method has a high accuracy in figuring out the defocus distance based on the single-shot color-multiplexed illumination image. We recorded the lens ring position during image acquisition and use the units of lens ring position to measure the amount of defocus. In order to better illustrate the accuracy of defocus prediction, we convert the lens ring position to Z position of objective lens through the calibrated proportional relationship shown in Fig. 1(e).

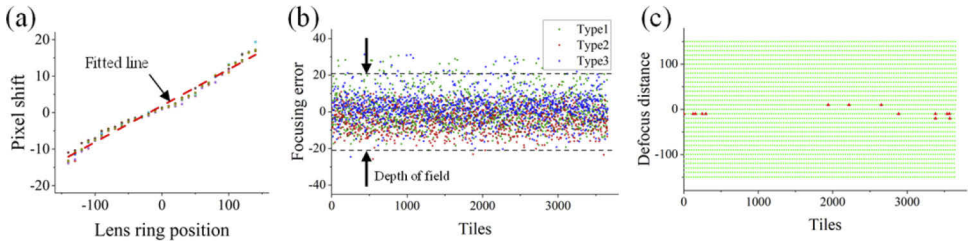


Fig. 7. Comparison of MI maximization and the proposed method. (a) Pixel shift calculated by MI maximization algorithm. (b) The focusing errors with the proposed method. (c) Defocus direction prediction with the proposed method.

In the actual image acquisition process, it is important to ensure the accuracy of the defocus direction prediction. If the defocus direction is predicted incorrectly, the quality of the captured image will be very poor. What is more serious, if the defocus distance exceeds the working range of focusing algorithm due to the incorrect prediction of the defocus direction, the system cannot complete the subsequent acquisition work. In our method, defocus direction is directly

Biomedical Optics EXPRESS

encoded in the acquired image under color-multiplexed illumination. Figure 7(c) shows the defocus direction judgement for test tiles, indicating that our method excels in defocus direction judgement with a single-shot image. The vertical axis of Fig. 7(c) is the actual defocus distance, which is within about ±10 μm defocus range (±160 lens ring position). The green dots indicate the correct defocus direction prediction tiles while the red triangles indicate the incorrect defocus direction prediction tiles. In the range out of the DOF, the proposed method predicts the defocus direction completely correct over the testing set. This is partly attributed to the color-multiplexed illumination method, which enables the images captured contain clear information about defocus direction.

3.3. Robustness to under-stained and thick samples

Another advantage of our method is the robustness against under-stained samples and thick samples. Typical FoVs of these two types are shown in Fig. 8(a) and Fig. 8(b). As shown in Fig. 8(c) for under-stained samples and Fig. 8(d) for thick samples, the pixel shift calculated by MI maximization algorithm deviates from the fitted curve shown in Fig. 7(a), indicating that the defocus distance cannot be figured out accurately. It may due to the insufficient lack and mutual influence of information between neighboring pixels for these two types of samples. The focusing errors of the proposed method are demonstrated in Fig. 8(e). The mean focusing errors are 0.34 μm and 0.33 μm, which are well fallen within the DOF, indicating the robustness of our method.

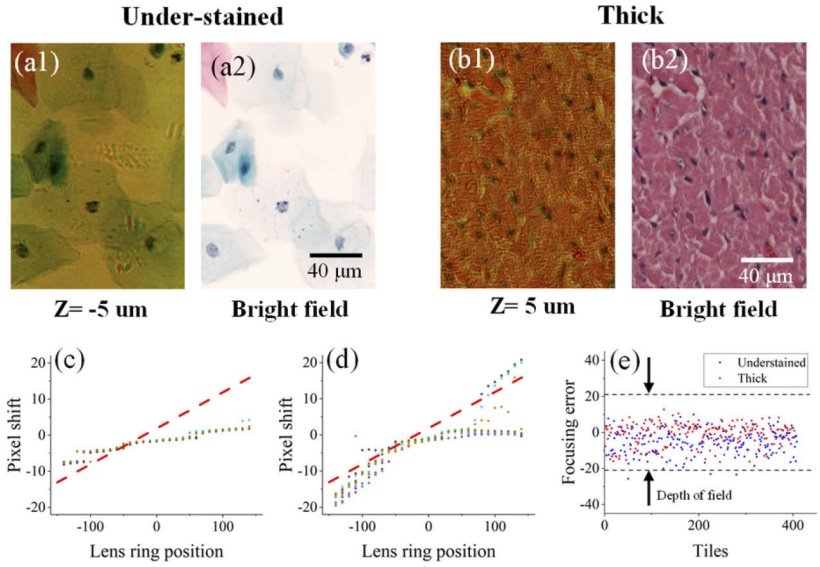


Fig. 8. Comparison of MI maximization algorithm and proposed method for under-stained samples and thick samples. (a) FoV of under-stained sample. (b) FoV of thick sample. (c) (d) Pixel shift calculated by mutual information maximization algorithm. (e) Focusing error with the proposed method.

3.4. Generalization ability

The generalization of neural networks has always been the focus of attention. In order to verify the generalization of the network trained in this paper, we tested the performance of our method using two sample types that have never appeared in the training set as a new test set and tested the performance of the proposed method.

Biomedical Optics EXPRESS

The focusing errors for different tiles are shown in Fig. 9(a), and the mean focusing errors are about 0.55 μm and 0.65 μm, respectively. The focusing error is still fallen within the DOF range, indicating that the trained model has good generalization ability. Typical FoV pictures of these two new types are shown in Fig. 9(b). The two types of samples have different texture and staining conditions, which further illustrates the robustness of our method.

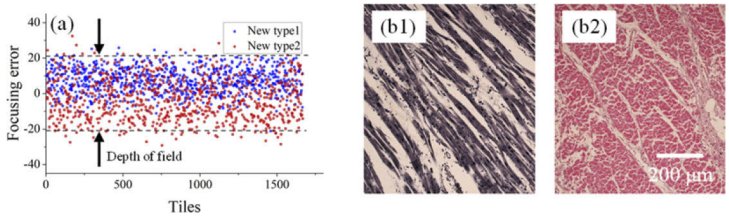
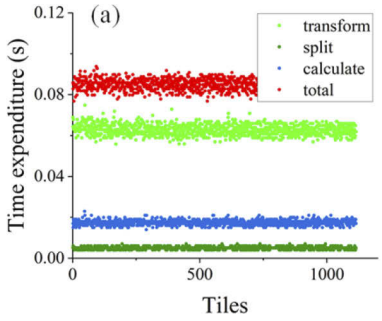


Fig. 9. (a) The focusing errors with the proposed method for two new types of pathological slides, which did not appear in the training set. (b) Autofocused images of the two new types of samples.

3.5. Time complexity analysis

In the process of acquiring a full-slice image, the time used for defocus calculation is an important indicator for evaluating a defocus calculation method. We tested time expenditure of our algorithm over a desktop with Intel i7-10700 processor, 32GB RAM, NVIDIA RTX2080Ti with 12G GPU memory. Results of processing time are shown in Fig. 10(a) and summarized in Fig. 10(b). In our setup, three steps are required to calculate the defocus distance of an image. First, the image needs to be converted into tensor format that can be calculated by the network and then is normalized. The time required for this process is 0.0628 ± 0.0026s. Then, the red and green channels of the image are divided into 12 non-overlapping sub-images. This step requires 0.0050 ± 0.0006s. Figuring out the defocus distance based on the 12 sub-images requires 0.0117 ± 0.0012s. Total time expenditure to calculate the defocus distance of an image is 0.085 ± 0.0027s. Obviously, the approach proposed can calculate the defocus distance within a short time, showing great application value. For the image acquisition of each FoV, we set a time delay of 0.3s for x-y stage movement, and 0.1s delay for focal plane adjustment. The total processing time for each FoV is about 0.5s.



(b)

Stage	Time expenditure(s)
Transform	0.0628±0.0026
Split	0.0050±0.0006
Calculate	0.0117±0.0012
Total	0.0850±0.0027

Fig. 10. Time expenditures to predict defocus distance.

3.6. Whole slide image capturing

Figure 11 shows a whole slide image captured with the reported system. The focus map (converted from lens ring position) is shown in Fig. 11(a), the depth information of each FoV is recorded

Biomedical Optics EXPRESS

during the image acquisition process. The whole slide image, as is shown in Fig. 11(b), is generated using the image stitching plug-in of imageJ. We apply linear blending in the overlapping regions. The acquisition time for this 1 cm by 1 cm sample image is ~56 s.

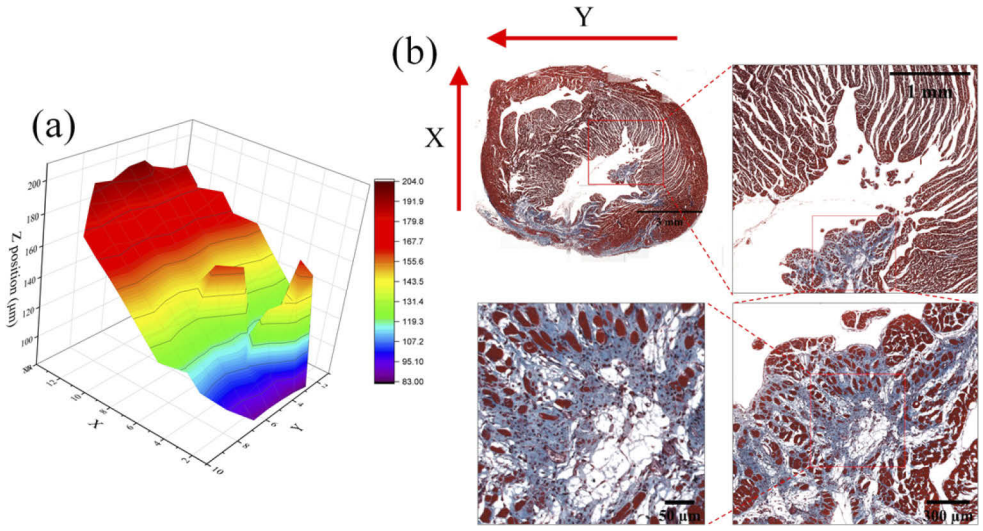


Fig. 11. (a) Focus map generated during image acquisition. (b) The captured whole slide image.

4. Conclusion

In summary, we reported a low-cost WSI scheme using deep learning based autofocusing. In hardware implementation, we use a customized adapter ring to connect the photographic lens and realizes precise control instead of using a precise mechanical stage for axial focusing adjustment. Compared with the invasive method to connect to the photographic lens, the proposed connection method provides a cheap and convenient solution for subsequent researches. We use a programmable LED array for sample illumination, which provides two illumination modes, brightfield illumination and partial coherent illumination. The adaption of our illumination equipment to traditional microscopy system is very convenient. In each FoV, we acquire a color-multiplexed illumination image for autofocusing before capturing the brightfield image. A neural network is further proposed to predict the defocus distance. Experimental result shows that the focusing error is well within the DOF, and our method is more robust against under-stained samples and thick samples. Moreover, the defocus distance can be accurately calculated for the sample types that have never appeared in the training set, indicating that our method has good generalization ability. Thanks to the 'light weight' network architecture, the processing time of autofocusing is extremely short. In the case of using GPU, the calculation time is less than 0.1 second, which shows the great potential for rapid high-throughput whole slide imaging.

Future work includes adding more types of samples in dataset to improve stability. We will also explore better network structures to further improve the accuracy of predictions and reduce the processing time.

Funding. National Natural Science Foundation of China (No. 61922048&62031023&61620106005&61827804); Science and Technology Planning Project of Shenzhen Municipality (No. JCYJ20200109142808034); Guangdong Special Support Plan (No. 2019TX05X187).

Disclosures. The authors declare no conflicts of interest.

Biomedical Optics EXPRESS

Data availability. 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.

References

1. R. Ferreira, B. Moon, J. Humphries, A. Sussman, and A. Demarzo, "The virtual microscope," *Proceedings: a conference of the American Medical Informatics Association /... AMIA Annual Fall Symposium. AMIA Fall Symposium* 4, 449–453 (1997).
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. 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).
4. 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).
5. Y. Liron, Y. Paran, N. G. Zatorsky, B. Geiger, and Z. Kam, "Laser autofocusing system for high-resolution cell biological imaging," J Microsc. **221**(2), 145–151 (2006).
6. Y. Zhang, H. Wang, Y. Wu, T. Miu, and O. Aydogan, "Edge sparsity criterion for robust holographic autofocusing," Opt. Lett. **42**(19), 3824 (2017).
7. U. R. R. T. Dong and J. Zeineh, "System and method for generating digital images of a microscope slide" (2005).
8. C. H. F. Velzel and P. F. Greve, "Apparatus for optically reading a record carrier and correcting focus error" (US, 1978).
9. K. Guo, J. Liao, Z. Bian, H. Xin, and G. Zheng, "InstantScope: a low-cost whole slide imaging system with instant focal plane detection," Biomed. Opt. Express **6**(9), 3210–3216 (2015).
10. J. Liao, L. Bian, Z. Bian, Z. Zhang, P. Charmi, H. Kazunori, Y. C. Eldar, and G. Zheng, "Single-frame rapid autofocusing for brightfield and fluorescence whole slide imaging," Biomed. Opt. Express 7(11), 4763 (2016).
11. 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, (2018).
12. J. Liao, S. Jiang, Z. Zhang, K. Guo, Z. Bian, Y. Jiang, J. Zhong, and G. Zheng, "Terapixel hyperspectral whole-slide imaging via slit-array detection and projection," J. Biomed. Opt. **23**(06), 1 (2018).
13. 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).
14. J. Liao, Z. Wang, Z. Zhang, Z. Bian, K. Guo, A. Nambiar, Y. Jiang, S. Jiang, J. Zhong, and M. Choma, "Dual-LED-based multichannel microscopy for whole-slide multiplane, multispectral, and phase imaging," J. Biophotonics **11**, 201700075 (2017).
15. C. Guo, Z. Bian, S. Jiang, M. Murphy, J. Zhu, R. Wang, P. Song, X. Shao, Y. Zhang, and G. Zheng, "OpenWSI: a low-cost, high-throughput whole slide imaging system via single-frame autofocusing and open-source hardware," Opt. Lett. **45**, 260 (2019).
16. S. Jiang, J. Liao, Z. Bian, K. Guo, and G. Zheng, "Transform- and multi-domain deep learning for single-frame rapid autofocusing in whole slide imaging," Biomed.. Opt. Express **9**(4), 1601 (2018).
17. T. Rai Dastidar and R. Ethirajan, "Whole slide imaging system using deep learning-based automated focusing," Biomed. Opt Express **11**(1), 480–491 (2020).
18. H. Pinkard, Z. Phillips, A. Babakhani, D. A. Fletcher, and L. Waller, "Deep learning for single-shot autofocus microscopy," Optica **6**(6), 794 (2019).
19. 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).
20. 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).
21. S. Jiang, Z. Bian, X. Huang, P. Song, H. Zhang, Y. Zhang, and G. Zheng, "Rapid and robust whole slide imaging based on LED-array illumination and color-multiplexed single-shot autofocusing," *Quantitative Imaging in Medicine and Survey* (2019).
22. J. F. Brenner, B. S. Dew, J. B. Horton, T. King, P. W. Neurath, and W. D. Selles, "An automated microscope for cytologic research a preliminary evaluation," J Histochem Cytochem. **24**(1), 100–111 (1976).
23. M. S. Andrew Howard, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam, "Searching for MobileNetV3," in *IEEE International Conference on Computer Vision* (2019).
24. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks" (2018).
25. J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), 7132–7141.