Research Article | Vol. 13, No. 1 / 1 Jan 2022 / Biomedical Optics Express 314
Biomedical Optics EXPRESS

Deep learning-based single-shot autofocus method for digital microscopy

JUN LIAO, XU CHEN,[1] GE DING,[1] PEI DONG,[1] HU YE,[1] HAN WANG,[1] YONGBING ZHANG,[2] AND JIANHUA YAO[1,*]

[1]Tencent AI Lab, Shenzhen 518054, China
[2]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen 518055, China
*jianhuayao@tencent.com

Abstract: Digital pathology is being transformed by artificial intelligence (AI)-based pathological diagnosis. One major challenge for correct AI diagnoses is to ensure the focus quality of captured images. Here, we propose a deep learning-based single-shot autofocus method for microscopy. We use a modified MobileNetV3, a lightweight network, to predict the defocus distance with a single-shot microscopy image acquired at an arbitrary image plane without secondary camera or additional optics. The defocus prediction takes only 9 ms with a focusing error of only ~1/15 depth of field. We also provide implementation examples for the augmented reality microscope and the whole slide imaging (WSI) system. Our proposed technique can perform real-time and accurate autofocus which will not only support pathologists in their daily work, but also provide potential applications in the life sciences, material research, and industrial automatic detection.

1. Introduction

In 2018, Google announced an augmented reality microscope (ARM) with real-time artificial intelligence integration for cancer diagnosis [1]. As the microscope is the most important tool for pathological diagnosis, the ARM has the potential to decrease the variability of pathological assessments and to alleviate the labor shortage of trained pathologists in regions such as rural areas [2]. However, defocus blur can greatly deteriorate the image quality and introduce tissue detail loss, thereby decreasing the reliability of the ARM or related AI-based microscopes. For ARM, defocus blur can occur due to the optical path length difference between the eyepiece ports and the camera port [3]. Pathologists are not trained to adjust the parfocal of the microscope and keeping the camera focused while reviewing the slides from the eyepiece simultaneously is difficult in practice.

US Food and Drug Administration (FDA) announced the approval of the first whole slide imaging (WSI) system for primary diagnosis in surgical pathology [4] in 2017. The WSI system has undergone an exponential period of growth for quantitative and streamlined slide reviewing [5]. We can regard the WSI system as a motorized high-capacity microscope with autofocus and auto-slide loading function. Although robust and high-throughput WSI systems are commercially available, their scanning speed is slow and their acquisition of well-focused digital slides remains inconsistent [6,7]. Pre-scanning a sample to acquire a focus map is the most adopted autofocus method for current WSI systems [8]. The focus map surveying requires z-stack acquisitions for focus plane estimation. Yet, the axial scanning of multiple images is time-consuming. Another issue is that skipping tiles can shorten the focus map surveying time at the cost of focus map accuracy [8].

In addition to the conventional time-consuming focus searching method through axial scanning, a variety of new autofocus methods for microscopy have emerged in recent years, which can be divided into three categories. The first category introduces additional illumination sources [9–13] and cameras [14–16] to the original microscope light path for defocus estimation. For

#446928 https://doi.org/10.1364/BOE.446928
Journal © 2022 Received 25 Oct 2021; revised 7 Dec 2021; accepted 7 Dec 2021; published 14 Dec 2021

Biomedical Optics EXPRESS

example, the Nikon Perfect Focus System introduces an additional infrared LED and a linear sensor to track the position of the slides [9]. This method adds cost and complexity to the system and only works for 2D thin slides. The second category is image-based and requires no additional optics but multiple shots for defocus estimation [17–19]. For example, Dastidar et al. use the difference of two shots at different focal planes and deep learning for defocus estimation [17]. The third category generates a virtual in-focus image according to the input blurry image using deep learning instead of estimating the defocus distance [20–24]. These methods require no additional hardware or multiple shots to refocus the image. For example, Wu et al. trained a network to virtually refocus a two-dimensional fluorescence image onto a user-defined focal plane within the sample [20]. Luo et al. use a deep learning-based offline autofocusing method that efficiently and blindly autofocus a single-shot microscopy image of a specimen that is captured at an arbitrary out-of-focus plane [24]. An image-generating approach would be more time-consuming as the image size increases. Also, there are always some doubts about virtually generated images and they may not be accepted for critical tasks.

In this paper, we demonstrate a deep learning-based single-shot autofocus method without any modifications to the original microscopy system for focal plane estimation. This approach only requires one image captured at an arbitrary plane by the inherent camera of the microscope to determine the focal plane. One can freely choose the motorized Z-stage, piezoelectric stage, or the tunable lens to finish the focus adjustment. Our novel method shows 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 distance. In the network's training phase, a motorized stage is used to collect z-stack images to train a modified MobileNetV3_small [25], a lightweight neural network, and achieves a 1/15 depth of field (DOF) focusing error of each 672*672 image patch. By measuring several image patches in one high-resolution image, the final focusing accuracy and robustness can be further improved.

We provide a specific autofocus implementation scheme on the ARM as an application example. Furthermore, we demonstrate our autofocus method can also be applied to focus map surveying for WSI system. We believe our work can help make AI the technological 'right hand' of pathologists. Our single-shot autofocus method is universal and can be applied to other imaging fields such as time-lapse live-cell imaging, and material research. Not limited to microscopy, it can also find future applications in industrial depth estimation and autopilot software.

2. Experiments and results

2.1. Single-shot autofocus for augmented reality microscopy

ARM overlays AI-generated information onto the current view of the sample in real-time, enabling seamless integration of AI into routine pathological workflows [1]. One of the serious problems encountered by this system in clinical trials is that defocused images lead to unreliable AI-diagnosis results, as an autofocus function is essential for ARMs to function. Another complaint is that an ARM can be too high for a comfortable sitting posture since the two parallel light paths (i.e., the image acquisition layer and AR projection layer) add to the height of the benchtop microscope as shown in Fig. 1(a).

Therefore, we aim to make improvements to the ARM to create a more practical system. First, we remove the image acquisition layer in the parallel light path and place the camera in the standard camera port on top of the microscope as shown in Fig. 1(b). This reduces the original height increase of the ARM by half. We also use two crossed polarizers to block all lights from the ARM screen from entering the camera. Second, we add an autofocus function using a deep learning network to estimate the defocus distance from a single image captured at an arbitrary image plane. Then, we use the liquid tunable lens to adjust the focus rapidly to complete the autofocus process. Our modified ARM is based on the Olympus BX43 microscope and the Lumenera Lt425 camera with a resolution of 2048*2048 pixels. The model of the tunable lens is

Biomedical Optics EXPRESS

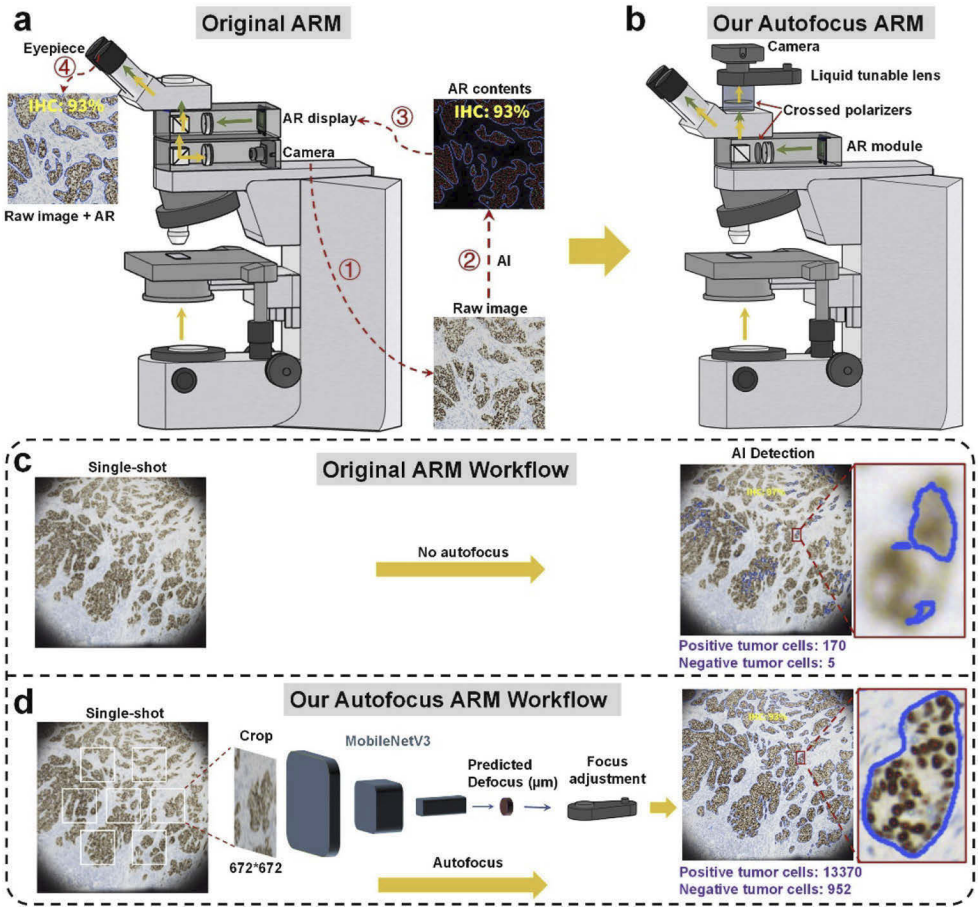

Fig. 1. Autofocus for ARM. (a) Setup scheme of the original ARM. The image acquisition layer and AR projection layer are added to a conventional microscope. First, the raw image captured by the camera is sent as input for AI processing. The outcome AR contents, for example, contours and text, are sent to the AR display. The user can observe the raw image overlaid with AR contents at the eyepiece port thanks to the beam splitters. (b) The setup scheme of our autofocus ARM. We install the camera and the tunable lens at the standard camera port of the conventional microscope. Only the AR projection layer is inserted into the infinity space. We use a pair of crossed polarizers to block the light of AR display from entering the camera. (c) Workflow of the original ARM: pathology AI algorithms are applied to the raw image without autofocus. (d) Workflow of our autofocus ARM: we first use a modified MobileNetV3 to estimate the defocus distance of the raw image. Then, we adjust the image focus through the liquid tunable lens. The in-focus image is then captured as the new input for pathology AI algorithms.

the Optotune EL-16-40-TC with a 16 mm aperture. A customized 0.4X adaptor connects the tunable lens and with the camera port. The augmented reality screen is the Sony ECX335S microdisplay. Figure 1(c) and (d) compares the workflow between the original ARM and our autofocus ARM. An immunohistochemistry (IHC) image captured by the original ARM at an arbitrary axial plane under a 10X/NA0.3 objective lens is directly sent for IHC AI processing. On the other hand, for our autofocus ARM, the captured image is first cropped to seven patches sized 672*672. We then predict the defocus distance of the seven patches respectively using a pre-trained defocus distance estimation network, a lightweight deep learning network modified

Biomedical Optics EXPRESS

from the MobileNetV3_small [25]. The tunable lens is responsible for focus adjustment according to the averaged predicted defocus distances. In Fig. 1(c) and (d), we also compare the IHC AI-detection [2] results of the view indicated by the red box for the image before autofocus and the autofocused image. In the raw image, only 170 positive tumor cells and five negative tumor cells are detected. In the autofocused image, 13,370 positive tumor cells and 952 negative tumor cells are detected. We present a partially enlarged image for a better visual perception. Please refer to Appendix B for more detail about our IHC AI algorithm.

As shown in Fig. 1(d), the defocus prediction network is based on MobileNetV3_small [25], a lightweight deep learning network, which is suitable for the real-time autofocus requirement of ARM. We make the below modifications to the original MobileNetV3_small: First, we change the input size to 672*672 from 224*224 to cover a larger field of view for reliable prediction. Second, we also change the classification output to regression output. In the training phase, taking the autofocus under a 10X objective lens as an example, we capture focal stacks of in-focus and defocus images using the HeidStar HDS-BFS-BX43-PRO-1, a motorized Olympus BX43 microscope equipped with a 10X/NA0.3 objective lens. The testing instrument or the autofocus ARM has the same setup as the training data collection instrument except for the stage of the latter is motorized. We capture 1500 z-stacks of pathological images, including IHC, Thinprep Cytology Test (TCT), as well as Hematoxylin and Eosin (H&E) slides in total. The same number of focus stacks are taken for each of the three pathological image types (IHC, TCT, HE). We divide the data into training, validation, and prediction sets at a ratio of 8:1:1, respectively. Each z-stack contains 25 images ranging from -36 µm to +36 µm with a step size of 3 µm. The "-" is facing away from the objective lens and "+" is facing the objective lens. Each raw image is cropped to seven 672*672 patches, which is the input size of the network. The axial step size, 3 µm, is not small enough compared to the 10 µm depth of field of the 10X/NA0.3 objective. To achieve better continuity of the defocus level when capturing z-stack images, we alternately capture -37 µm to +35 µm, -36 µm to +36 µm, -35 µm to +37 µm. To label each image with its defocus distance, we use a Brenner Gradient [26,27] to locate the focal plane with subpixel resolution. The label value is the ground truth of the defocus distance for the image.

We train four autofocus networks in total, one model each trained for IHC, TCT, and H&E, and one mixed model trained with all mixed data. As shown in Fig. 2(a)-(c), the focusing error of the IHC, TCT, and H&E models are 0.82 ± 0.61 µm, 0.73 ± 0.75 µm, and 0.73 ± 0.61 µm respectively. The mixed model provides the best autofocus performance with a focus error of 0.68 ± 0.58 µm, as shown in Fig. 2(d). The depth of field of the 10X/NA0.3 objective lens is 10 µm.

Figure 2(a-d) indicates that not only can we distinguish different degrees of defocus but also positive defocus from negative defocus. We will further discuss the mechanism of distinguishing positive and negative defocus with a single shot in the Discussion section.

We choose the MobileNetV3-small after careful comparison. As shown in Fig. 2(e), we compare the following eight popular deep learning networks, including deep and heavy networks as well as lightweight networks using H&E data: MobileNetV3-small, MobileNetV3-large [25], ResNet50 with a ghost module [28,29], DenseNet121 [30], ResNet50 [28], MobileNetV3-small with a ghost module [29], MobileNetV3-small with an FFT module [31], and ShuffleNetV2 [32]. MobileNetV3-small has just half number of parameters of MobileNetV3-large. In the end, MobileNetV3-small returns the best results.

The above deep learning networks we choose are among the most cited networks on classification, object detection and semantic segmentation in recent years. We can divide these networks into two types: heavy networks such as the ResNet50 and DenseNet121 and lightweight networks such as the MobileNetV3 and ShuffleNetV2. The ghost module and FFT module are integrated into the architecture of the ResNet50 and MobileNetV3 with verified performance improvement [29,31]. The same training data (100 z-stacks of HE images) and same training conditions (100 training epochs, Adam optimizer, multistep learning strategy and smooth L1 loss function, etc.)

Biomedical Optics EXPRESS

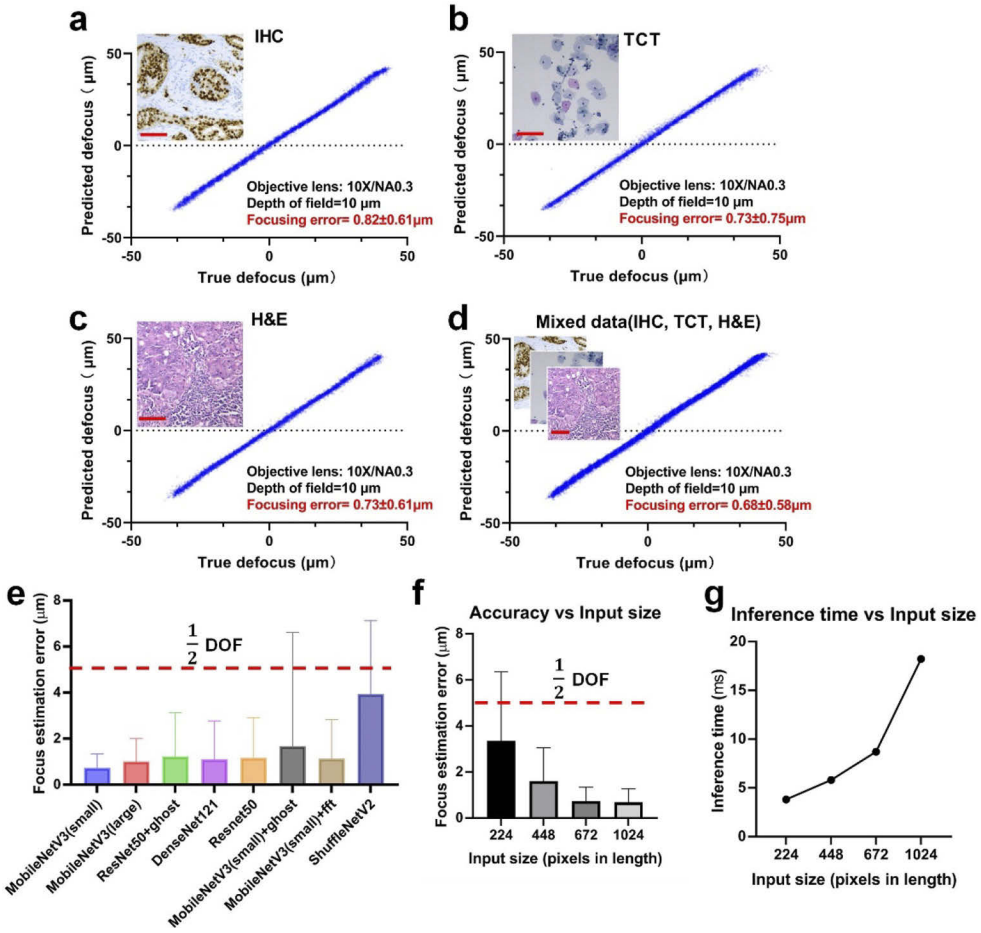

Fig. 2. Autofocus performance of the modified MobileNetV3 small on test data. (a)-(d) Scatter plots show the testing performance of the autofocus models trained by IHC, TCT, H&E and mixed data separately. (e) Autofocus performance comparison of eight selected deep learning networks. (f) Different input image size lead to different autofocus performance. (g) Inference time versus input image size. The red scale bars in (a)-(d) indicate 100 μm.

are configured to guarantee a relatively fair comparison. Figure 2(e) shows that most of the testing networks demonstrate good defocus estimation ability after the training (mean error < half DOF). The MobileNetV3_small is the one which masters the defocus estimation ability. We believe that this benefits from the advanced design of MobileNetV3: the optimal number of convolution kernels and channels obtained using the NetAdapt, inherited depth separable convolution and residual structure with linear bottleneck from V1 and V2, and the newly introduced activation function hard-swish which is verified with the ability to effectively improve the accuracy of the network [25].

Figure 2(f) shows that a larger input image size gives a more accurate output. However, the inference time increases exponentially as the input image size increase as shown in Fig. 2(g). We choose input size as 672*672 as a tradeoff between accuracy and efficiency.

The autofocus procedure on an ARM takes 59 ms in total: 25 ms for capturing an image and removing noise with a 3*3 median filter, 9 ms for defocus estimation, and 25 ms for focus adjustment with the tunable lens. The testing computer runs on Linux and has an Intel(R)

Biomedical Optics EXPRESS

Xeon(R) CPU E5-2680 v4 @ 2.40 GHz. One Tesla P40 GPU (24G memory) is assigned to the autofocus module.

2.2. WSI focus map surveying using our single-shot autofocus method

WSI systems automatically image the whole slide, turning a physical slide into a digital one. This enables doctors to stay away from the microscope to conduct remote pathological diagnoses and consultations. The most adopted autofocus method for WSI system is to acquire a focus map in the beginning. As shown in Fig. 3(a), an external camera captures a scout image of the slide. Then, focus points spacing over the entire sample excluding the background are selected automatically. At each focus point, conventionally, the system will capture a z-stack image to find the best focal plane. By interpolating the coarse focus map, a full focus map that guides the scanning for sharp WSI output can be obtained.

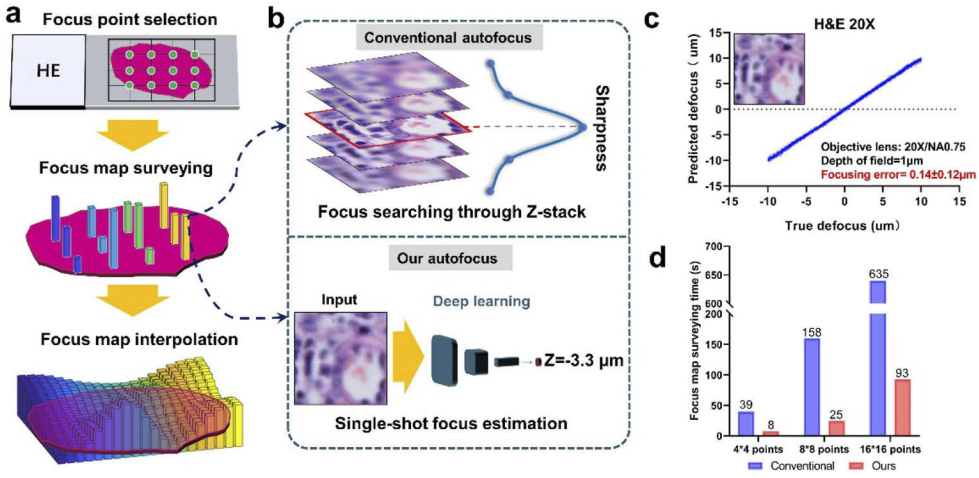

Fig. 3. Comparison between conventional and our focus map surveying method on WSI system. (a) Workflow of focus map surveying method. (b) Comparison of conventional defocus estimation method (top) by acquiring a z-stack and our single-shot deep learning approach (bottom). (c) The scatter plot shows the testing results of autofocus performance of our approach on a WSI system equipped with a 20X/NA0.75 objective lens. (d) Time-consumption comparison between conventional focus map surveying method and ours.

The proposed single-shot autofocus network can be implemented on the WSI system as well, making the autofocus process of WSI much more efficient. Since WSI system already has a built-in focus adjustment module, such as the motorized Z-stage or piezoelectric stage, no liquid tunable lens is required (unlike the previously mentioned autofocus ARM). Hence, we only need to introduce the defocus prediction network for WSI autofocus.

The conventional autofocus method, as shown in Fig. 3(b), locates the focal plane within a z-stack image by evaluating the image's sharpness. However, scanning the z-stack images is very time-consuming since it requires multiple (e.g. 10) axial mechanic moves and image acquisitions. We use the proposed single-shot autofocus method for focus map surveying. High accuracy of defocus distance estimation can be achieved in just 34 ms for each focus point, at least 25 times faster than conventional methods. We retrain the autofocus network for WSI system equipped with a 20X/NA0.75 objective lens by capturing 1,000 H&E z-stacks for model training. Each z-stack ranges from -10 μm to 10 μm with a step size of 0.5 μm. We crop out nine 672*672 patches from each raw image and use image rotation plus image flipping for data augmentation. The model structure is the same as the modified version for our autofocus ARM. We use another

Biomedical Optics EXPRESS

300 z-stacks for testing, as shown in Fig. 3(c), and get a focus error of 0.14μm±0.12 μm. The depth of field of the objective is 1 μm.

Figure 3(d) shows the comparison of the time consumed to obtain the focus map under different amounts of focus points between the conventional z-stack autofocus method and ours. It indicates that as the focus points of the focus map increase, the sampling of focus map will be more refined, but the time taken will increase significantly. Our method needs less than 19% of conventional focus map surveying time and spends most of the time (~85%) on X-Y movement.

Figure 4 shows a real example of a focus map acquired by our autofocus method. Figure 4(a) is the thumbnail image of a Lymph node H&E sample. We acquire the focus map of a 10mm*10 mm area as indicated by the red box in Fig. 4(a). For comparison, we survey two 16*16-points focus maps using our method and conventional method respectively. The results are shown in Fig. 4(b)

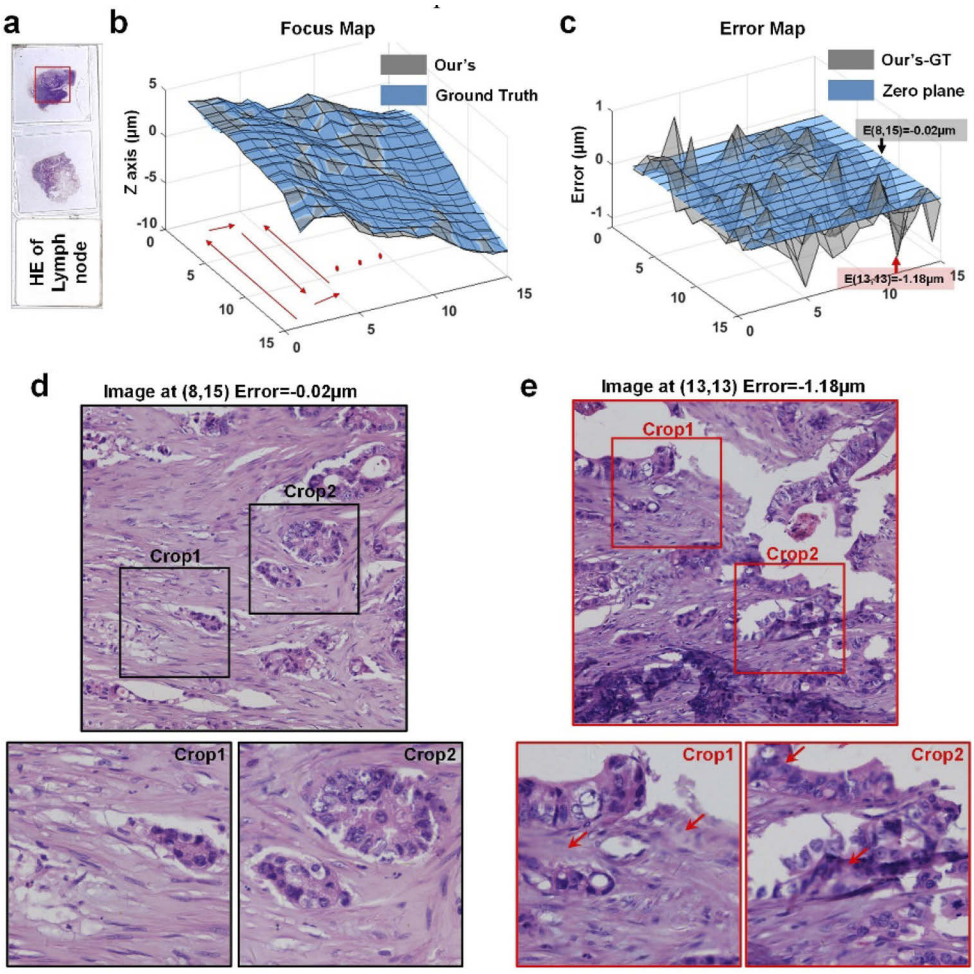

Fig. 4. Actual focus map surveying example using our single-shot autofocus approach. (a) Thumbnail image of the Lymph node H&E slide. The red box indicates the focus map surveying area. (b) Focus map acquired using our deep learning autofocus method. (c) Error map of the focus map using the Brenner Gradient-based z-stack searching method as ground truth. (d) Scanned image using our autofocus approach at the point indicated with a black triangle in c. (e) Scanned image using our autofocus approach at the point indicated with a red triangle in c. Red arrows in Crop1 and Crop2 indicate areas at different focal planes.

Biomedical Optics EXPRESS

with grey color indicating our method and blue color indicating conventional method or ground truth. We use the scanning trajectory indicated by red arrows in Fig. 4(b) to ensure that every two adjacent scanning points are also adjacent in the actual spatial position during the scanning process. Since the difference of the defocus distance of adjacent points will not be too long, compared to the Zigzag scanning trajectory, our trajectory can prevent the defocus of the next focus point from exceeding the prediction range of the defocus prediction network when changing lines. Figure 4(c) shows the error map, which is the difference between the focus map obtained by our method and the ground truth. The mean error of the focus map acquired with our approach is 0.28μm±0.32 μm. The depth of field of the 20X/NA0.75 objective lens used is 1 μm. From the error map, most focus points have a focus error within 0.5 μm such as the black arrow indicated points (x=8,y=15). The red arrow indicated point (x=13,y=13) in Fig. 4(c) has a focus error of -1.18 μm, which is larger than the depth of field. According to our focus map, we scan the entire slide to acquire the whole slide image. We show the scanned image of the points at the black and red arrows pointed regions on Fig. 4(c) in Figs. 4(d) and (e), respectively. Figure 4(d) shows a typical successful case of autofocus with a focus prediction error of -0.02 μm. We can observe that the cropped images are successfully focused. In Fig. 4(e), we show a typical "failure" case with a focusing error of -1.18 μm. However, we can find that this image contains many thick areas (pointed by red arrows in Crop1) and folded areas (pointed by red arrows in Crop2) in Fig. 4(e). Crop1 and Crop2 both have a size of 672*672. We argue that in this case, the ground truth calculated by the conventional z-stack searching does not serve as a universal standard solution. We recommend an axial scanning near this plane to get a composite all-in-focus image. The advantage of our patch sampling approach is that the predicted defocus distances of the seven sub-field-of-views (672*672) can tell us the focus distribution and variation of the original image (2K*2 K). We have two strategies when handling different samples including flat, uneven, and tilted ones. First, if the variation of the predicted defocus distances is smaller than the depth of field of the objective lens. We consider the field of view is even and suggest using the averaged

Table 1. Comparison of the state-of-the-art autofocus methods for microscopy

Method	Strategy	Deployment Complexity	Optical System	Focusing Error(um)	Focusing Error/DOF
Pinkard [10]	Modified illumination (extra LEDs) + Single-shot **deep learning** (FFT as input)	★★★★	20X/NA0.5 DOF=2.6μm	1.38	5.3/10
Jiang [18]	Modified illumination (extral LEDs) + Single shot **deep learning** (three-domain-image as input)	★★★★	20X/NA0.75 DOF=1μm	0.21 + 0.17	2.1/10
Dastidar [17]	Two shots at different axial planes as **deep learning** inputs	★★	20X/NA0.75 DOF=1μm	0.19 + 0.18	1.9/10
Zhang [13]	Separate autofocus module based on laser triangulation (extra laser illumination + secondary camera).	★★★★	50X/NA0.55 DOF=1.96μm	0.2	1.0/10
Liao [11]	Modified illumination (extra LEDs) + autocorrelation analysis.	★★★★	20X/NA0.75 DOF=1μm	0.08	**0.8/10**
Ours	Single shot at arbitrary axial plane as **deep learning** input	★	**20X**/NA0.75 DOF=1μm	0.14 + 0.12	**1.4/10**
			10X/NA0.3 DOF=10μm	0.68 + 0.58	**0.68/10**

Biomedical Optics EXPRESS

defocus distance to guide the focus adjustment. Second, if the variance of the predicted defocus distances is larger than the depth of field of the objective lens. We consider the field of view is uneven and suggest an axial scanning according to the variation range of the predicted defocus distances.

We compare our autofocus method with the state-of-the-art autofocus method for microscopy in Table 1 including deep learning-based and non-deep learning-based methods. DOF stands for depth of field of the objective lens. Since these approaches don't have exactly the same experiment setup (e.g. magnification, NA, DOF), we use the ratio of focusing error to DOF for a relatively fair comparison. From Table 1, our single-shot autofocus method has the highest focusing accuracy among the deep-learning-based methods and requires the least modification to the conventional microscope.

3. Discussion

In the scatter plots shown in Fig. 2 and Fig. 3, our single-shot deep learning autofocus method distinguishes the positive defocus and the negative defocus of the sample very well. However, when we use blur kernels (e.g. Gaussian blur) to simulate an out-of-focus image, there is no difference between the images on both sides of the focal plane, making it impossible to distinguish the defocus direction. The real-world defocus, which contains axial asymmetric spherical aberration and chromatic aberration, is more complicated than these common simulation methods. Based on our results, we deduce that the asymmetry allows distinguishing the focus direction of the real-world defocus image. To evaluate the asymmetry, we simulated the point spread functions (PSF) of different wavelengths at different focal planes in Fig. 5(a) with Zemax software by ray tracing. The objective lens is 10X/NA0.3.

As shown in Fig. 5(a), the PSFs are asymmetric on both sides of the focal plane. On the other hand, at the same focal plane, the PSFs of different wavelengths are also different. And this is where chromatic aberration comes from. We believe the asymmetric spherical and chromatic aberration is detectable by regular cameras (e.g. 5.5 µm pixel size), hence making the defocus directions distinguishable. In Fig. 5(b), we show the H&E images at different focal planes under a 10X/NA0.3 objective lens. We can observe color differences, especially at the white blank areas, at -20 µm and 20 µm defocus planes.

There are many autofocus techniques for microscopy, such as using additional autofocus illumination optics for defocus distance estimation [9–12]. However, many autofocus methods are not suitable for ARM. For example, Pinkard et al. used one or a few off-axis LEDs to guide defocus distance prediction with a single-shot using deep learning [10]. This is suitable for high NA objective lenses, such as autofocus for WSI system. As the NA shrink for low magnification lens such as 10X and 4X, the room left for off-axis LED is very small. Also, the focusing accuracy will decrease as the NA shrinks. However, for ARM, due to the depth-of-focus difference between eyepiece and camera port, 4X and 10X are the applications that need autofocus most. And additional illumination sources are not readily available as plug-and-play modules for current microscopes used in pathology. Tathagato [17] proposed using the difference image of two-shot for defocus distance prediction using deep learning. The required multiple shots at different focal planes are not efficient for the ARM, which has a strong requirement for real-time output. In this paper, we propose a single-shot autofocus method for ARM using a lightweight deep learning network without introducing additional illumination sources or cameras. We install the liquid lens before the camera to solve the parfocal problem. We do not choose the motorized z stage or install focus adjustment hardware connected with the objective lens [33,34] to adjust the lens focus since this does not decouple the autofocus for the user from the camera. The liquid lens is not the only option for rapid focus adjustment. Alternatively, a piezoelectric stage in front of the camera or an autofocus camera with a built-in electric stage for sensor axial movement are also efficient choices for fast autofocus. Shortening the optical stack of the original ARM does not

Biomedical Optics EXPRESS

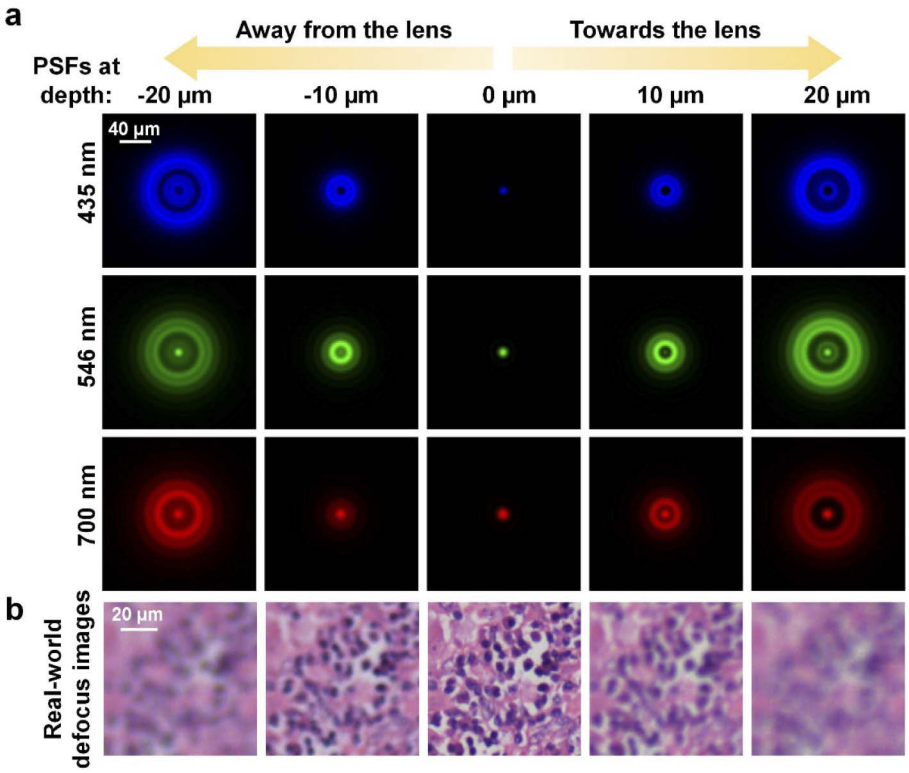

Fig. 5. Understand the single-shot autofocus. (a) shows Zemax simulated PSFs of different wavelengths at different focal planes. (b) shows the real-world defocus images at corresponding focal planes as (a). The objective lens is 10X/NA0.3.

impact the autofocus performance. This design change helps make more room for the installation of the liquid lens or other autofocus tools. Another advantage is that shortening the stack is more user-friendly for pathologists by reducing the height of the microscope.

For focus map surveying of WSI system, the conventional method conducts axial scanning to find the focus, which is not efficient. However, our deep learning method requires only a single shot to predict the defocus distance of the current field of view. One can crop more than seven images (the default setup in this study) from a raw image to get more robust prediction results, and parallel computing will make the total prediction time almost unchanged.

Compared with the state-of-the-art autofocus methods (Table 1), our autofocus method demonstrates a new idea to estimate the defocus distance: using the data-driven method to decode the defocus information from the captured raw image itself. In contrast, conventional methods use additional hardware such as extra illuminations or multiple shots to modulate the axial defocus information to image planes. The advantages of our autofocus method come from at least four aspects: First, we collect large training data (500 z-stacks each for HE, IHC, and TCT). Second, instead of using the original image for focus prediction, we use the patch sampling approach to divide the field of view of the captured image, hence getting a finer and more accurate focus (eg. a tilted sample). We also update the defocus label for each patch to compensate for focus variations at different patches. Third, the two strategies (single focus adjustment or axial scan) we used when handling uneven focus. Fourth, the efficient network structure of the MobileNetV3_small.Our autofocus method still shows room for improvement due to its novelty. One limitation is that we do not yet consider motion blur which is caused by

Biomedical Optics EXPRESS

fast stage movement. The ability to predict the defocus distance of images with motion blur can avoid interrupting pathologists' slide reading process on the ARM. For WSI system, the ability to predict the defocus distance of images with motion blur can also improve the focus map surveying efficiency hence shortening the time of the pathological diagnosis cycle.

The single-shot autofocus method is universal and should not be limited to the ARM and WSI system. For live-cell imaging or time-lapse imaging, a popular autofocus platform is the Nikon Perfect Focus System [9] which performs autofocus with a reference infrared beam to track the slide surface's fluctuation. The drawback of the Nikon perfect focus system and related autofocus techniques is that if the live cell or other moving target grows or moves above or below the reference plane, the autofocus will fail. Since our autofocus method is image-based, we can perform consistent autofocus for our interested target such as the live-cell which is easy to be located and separated from the background.

In summary, we report a single-shot autofocus method using a lightweight deep learning network for microscopy. We also propose a specific scheme of autofocus ARM using the autofocus network and the liquid tunable lens. We also incorporate this autofocus method into WSI system for focus map surveying. Compared with the state-of-the-art deep learning-based autofocus method, our approach is significantly more accurate and easier to deploy. Hence, our work will allow ARM and related AI products to enter the pathology department to support the limited pathologist workforce. We believe our paper provides a new idea for autofocus not just limited to ARM and WSI systems but can also find use in life science imaging, photography, and industrial machine vision.

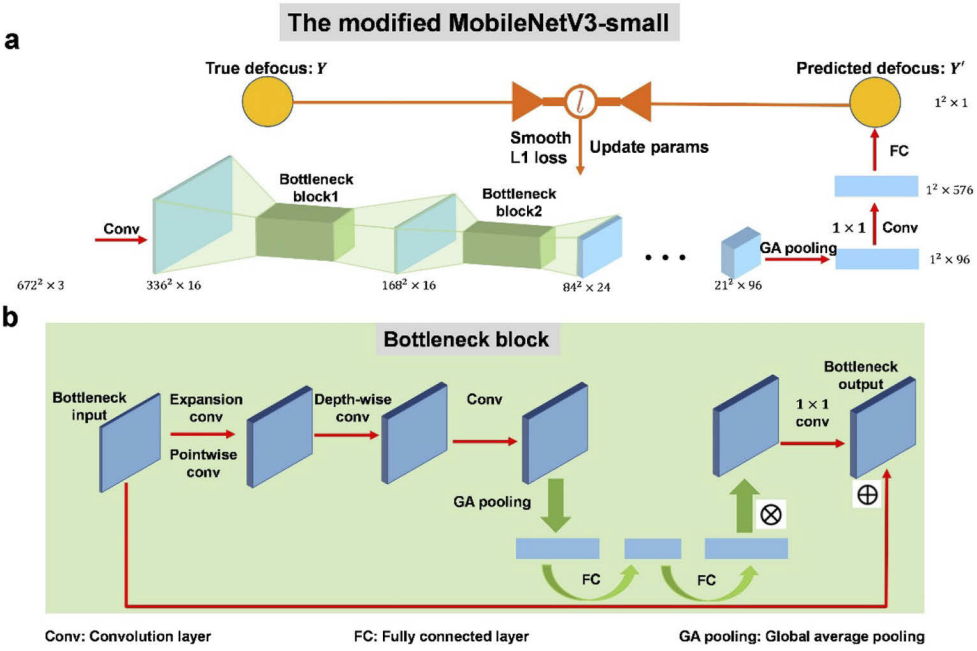

Fig. 6. The architecture of our single-shot autofocus deep learning network. (a) shows the single-shot autofocus network which is modified from the MobileNetV3-small. The network inputs an RGB image patch captured at an arbitrary focal plane and outputs the predicted defocus distance. (b) shows the inside of bottleneck blocks in a.

Biomedical Optics EXPRESS

Appendix A: Architecture of the single-shot autofocus network

Figure 6 shows our single-shot autofocus network which is modified from the MobileNetV3-small [25]. First, we change the input size to 672*672 from 224*224 to cover a larger field of view for reliable prediction. Second, we change the classification output to regression output. We adopt the Adam optimizer and smooth L1-loss for the autofocus network. To get maximal autofocus performance, we use population-based training(PBT) [35], a hyperparameter optimization method, to search the best dropout, learning rate, momentum, weight_decay, width_mult, etc.

Appendix B: IHC AI algorithm

The IHC AI algorithm [2] encodes contextual features from multiple image levels via convolutional neural networks (ResNet-18 backbone) and produces probability maps by integrating upsampled feature maps towards the final detection of cell centroids and segmentation of tumors [36,37]. Receptive fields of different sizes are employed to detect nucleus centroids and delineate tumor regions since the former focuses more on relatively local information while the latter proves to benefit from grasping the global view of the entire IHC image patch.

Appendix C: Brenner gradient

We use the Brenner gradient [26,27], an efficient and robust edge detector, as the figure of merit to evaluate the sharpness of the image. The Brenner gradient value of an image is calculated as follows:

$$B = \sum_{i=1}^{N} \sum_{j=1}^{M} [s(i,j) - s(i+2,j)]^2$$

This algorithm computes the first difference between a pixel and its neighbor with a distance of 2. s(i, j) is the pixel value at (i, j) coordinate position. N and M represent the number of pixels in the i and j directions. B is the final Brenner gradient value.

Appendix D: Calibration curve of liquid lens focus power in current versus defocus distance

We install the camera and tunable lens as shown in Fig. 1(a) on the motorized HDS motorized microscope (based on the BX43 microscope with 10X/NA0.3). The parfocal is adjusted when the current of the liquid lens is set to 0 mA. Then, we set the current of the tunable lens to -120 mA and adjust the motorized Z-stage to find the focal plane with the Brenner gradient, as the first point shown in Fig. 7. We get the whole calibration curve in the same manner with a current step size of 10 mA.

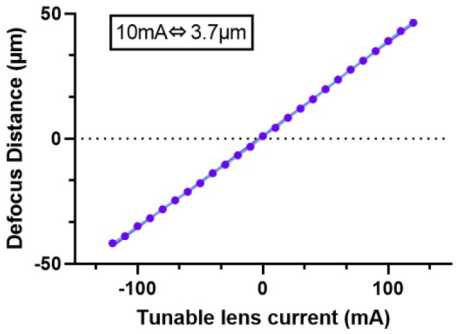

Fig. 7. The calibration curve of liquid lens focus power in current versus defocus distance under 10X/NA objective lens.

Biomedical Optics EXPRESS

Funding. Tencent AI Lab.

Disclosures. The authors declare no conflicts of interest.

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. P. H. C. Chen, K. Gadepalli, R. MacDonald, Y. Liu, S. Kadowaki, K. Nagpal, T. Kohlberger, J. Dean, G. S. Corrado, J. D. Hipp, C. H. Mermel, and M. C. Stumpe, "An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis," Nat. Med. 25(9), 1453–1457 (2019).
2. L. Cai, K. Yan, H. Bu, M. Yue, P. Dong, X. Wang, L. Li, K. Tian, H. Shen, J. Zhang, J. Shang, S. Niu, D. Han, C. Ren, J. Huang, X. Han, J. Yao, and Y. Liu, "Improving Ki67 assessment concordance with ai-empowered microscope: a multi-institutional ring study," Histopathology 79(4), 544–555 (2021).
3. 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), 1–21 (2020).
4. "FDA allows marketing of first whole slide imaging system for digital pathology | FDA," https://www.fda.gov/news-events/press-announcements/fda-allows-marketing-first-whole-slide-imaging-system-digital-pathology.
5. F. Ghaznavi, A. Evans, A. Madabhushi, and M. Feldman, "Digital imaging in pathology: whole-slide imaging and beyond," Annu. Rev. Pathol. Mech. Dis. **8**(1), 331–359 (2013).
6. J. R. Gilbertson, J. Ho, L. Anthony, D. M. Jukic, Y. Yagi, and A. V. Parwani, "Primary histologic diagnosis using automated whole slide imaging: a validation study," BMC Clin Pathol **6**(1), 4 (2006).
7. J. Ho, A. V. Parwani, D. M. Jukic, Y. Yagi, L. Anthony, and J. R. Gilbertson, "Use of whole slide imaging in surgical pathology quality assurance: design and pilot validation studies," Hum. Pathol. **37**(3), 322–331 (2006).
8. M. Montalto, R. McKay, and V. Baxi, "The accuracy of dynamic predictive autofocusing for whole slide imaging," J Pathol Inform **2**(1), 38 (2011).
9. J. Silfies, E. Lieser, S. Stanley, and M. Davidson, "The Nikon Perfect Focus System (PFS) | Nikon's MicroscopyU," https://www.microscopyu.com/tutorials/the-nikon-perfect-focus-system-pfs.
10. H. Pinkard, Z. Phillips, A. Babakhani, D. A. Fletcher, and L. Waller, "Deep learning for single-shot autofocus microscopy," Optica **6**(6), 794 (2019).
11. J. Liao, Y. Jiang, Z. Bian, B. Mahrou, A. Nambiar, A. W. Magsam, K. Guo, S. Wang, Y. Cho, and G. Zheng, "Rapid focus map surveying for whole slide imaging with continuous sample motion," Opt. Lett. **42**(17), 3379–3382 (2017).
12. 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," Quant. Imaging Med. Surg. **9**(5), 823–831 (2019).
13. X. Zhang, F. Fan, M. Gheisari, and G. Srivastava, "A novel auto-focus method for image processing using laser triangulation," IEEE Access **7**, 64837–64843 (2019).
14. M. Montalto, R. Filkins, and R. McKay, "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).
15. M. E. Bravo-Zanoguera, C. A. Laris, L. K. Nguyen, M. Oliva, and J. H. Price, "Dynamic autofocus for continuous-scanning time-delay-and-integration image acquisition in automated microscopy," J. Biomed. Opt. **12**(3), 034011 (2007).
16. 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 (2016).
17. T. Rai Dastidar and R. Ethirajan, "Whole slide imaging system using deep learning-based automated focusing," Biomed. Opt. Express **11**(1), 480 (2020).
18. S. Jiang, J. Liao, Z. Bian, K. Guo, Y. Zhang, and G. Zheng, "Transform- and multi-domain deep learning for single-frame rapid autofocusing in whole slide imaging," Biomed. Opt. Express **9**(4), 1601 (2018).
19. A. Shajkofci and M. Liebling, "Spatially-variant CNN-based point spread function estimation for blind deconvolution and depth estimation in optical microscopy," IEEE Trans. on Image Process. **29**, 5848–5861 (2020).
20. 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).
21. Q. Li, X. Liu, K. Han, C. Guo, X. Ji, and X. Wu, "Rapid whole slide imaging via learning-based two-shot virtual autofocusing," arXiv (2020).
22. C. Jiang, J. Liao, P. Dong, Z. Ma, D. Cai, G. Zheng, Y. Liu, H. Bu, and J. Yao, "Blind deblurring for microscopic pathology images using deep learning networks," (2020).
23. L. Jin, Y. Tang, Y. Wu, J. B. Coole, M. T. Tan, X. Zhao, H. Badaoui, J. T. Robinson, M. D. Williams, A. M. Gillenwater, R. R. Richards-Kortum, and A. Veeraraghavan, "Deep learning extended depth-of-field microscope for fast and slide-free histology," Proc. Natl. Acad. Sci. U. S. A. **117**(52), 33051–33060 (2020).
24. 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).
25. A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for MobileNetV3," *Proc. IEEE Int. Conf. Comput. Vis. 2019-October*, 1314–1324 (2019).

Biomedical Optics EXPRESS

26. 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).
27. 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 (2008).
28. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2016 (IEEE Computer Society, 2016), pp. 770–778.
29. K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, *GhostNet: More Features from Cheap Operations* (n.d.).
30. G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *Proc. 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017* (2016), pp. 2261–2269.
31. C. Qiao, D. Li, Y. Guo, C. Liu, T. Jiang, Q. Dai, and D. Li, "Evaluation and development of deep neural networks for image super-resolution in optical microscopy," Nat. Methods **18**(2), 194–202 (2021).
32. N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: practical guidelines for efficient CNN architecture design," Lect. Notes Comput. Sci. **11218**, 122–138 (2018).
33. M. Bathe-Peters, P. Annibale, and M. J. Lohse, "All-optical microscope autofocus based on an electrically tunable lens and a totally internally reflected IR laser," Opt. Express **26**(3), 2359 (2018).
34. *Fast Laser Autofocus for Microscopes Pifoc Piezo Nano Drives, Semicon, Biotech, High-Resolution Microscopy, Microscope Stage, Microscope Positioner* (n.d.).
35. M. Jaderberg, V. Dalibard, S. Osindero, W. M. Czarnecki, J. Donahue, A. Razavi, O. Vinyals, T. Green, I. Dunning, K. Simonyan, C. Fernando, and K. Kavukcuoglu, "Population based training of neural networks," arXiv:1711.09846 (2017).
36. A. Chaurasia and E. Culurciello, "LinkNet: exploiting encoder representations for efficient semantic segmentation," *2017 IEEE Vis. Commun. Image Process. VCIP 2017 2018-January*, 1–4 (2017).
37. L. Yang, Y. Zhang, J. Chen, S. Zhang, and D. Z. Chen, "Suggestive annotation: a deep active learning framework for biomedical image segmentation," Lect. Notes Comput. Sci. **10435**, 399–407 (2017).