Automated focus distance estimation for digital microscopy using deep convolutional neural networks

Tathagato Rai Dastidar
SigTuple Technologies,
Bengaluru, KA 560102, India
trd@sigtuple.com

Abstract

An essential component of an automated digital microscopy system is auto focusing, which involves moving the microscope stage along the vertical axis to find the position where the underlying image is the sharpest. Auto focusing algorithms deployed in current commercially available digital microscopes often cannot match the efficiency of a trained human operator. Traditionally, auto focusing has been achieved by acquiring multiple images in the vertical direction and maximising a measure of image sharpness. This paper presents a method for auto focusing based on deep convolutional neural networks (CNN). Given two images in the vertical focus stack, the CNN predicts the optimal distance the stage needs to be moved to achieve best focus, relative to the current position. The method was trained and results are demonstrated on a publicly available data set. It is shown to outperform previously published work on this data set. The compute and memory requirements of the model are shown to be ideal for deployment in an edge device with limited computing resources.

1. Introduction

Manual microscopic review of biological samples remains the gold standard for diagnosis of diseases in several types of sample. This includes analysis of tissues (histopathology), blood (haematology), study of micro organisms (microbiology), and many more. Automated digital microscopes, also known as whole slide scanner (WSI) systems, aim to partially automate this process [1]. They scan and digitise the physical sample to create what is commonly known as a "virtual slide". This virtual slide can then be viewed on a computer screen and analysed remotely. Multiple medical experts can collaboratively analyse them. They can be preserved well beyond the shelf life of the physical sample. The virtual slide images can be further screened or analysed by artificial intelligence (AI) based systems, such as [5, 13], to detect various abnormalities. Recently, a WSI system has also been approved for primary diagnostic use by the US FDA [1].

An essential component of any automated microscope is the auto focus system, which brings the sample at the optimal position in the vertical axis for it to be imaged without defocus artefacts. Auto focusing for digital microscopy has been an active area of research for more than 4 decades [3]. However, in spite of advances in image capture speed and digital compute speed, automated microscopes still cannot match the performance of a skilled human operator while focusing on a given field of view, or keeping the sample in focus during observation [20, 9]. Consequently, digital slide scanning is time consuming. Though it is gaining popularity in the field of histopathology, digital slide scanning is yet to find its use for high volume laboratory tests such as peripheral blood smear analysis or urine microscopy. Thus, an intelligent and efficient automated microscopy system which can match or exceed human efficiency can greatly further the cause of digital pathology. It can enable telepathology, which till now is yet to see widespread adoption. It will also help enable rapid data creation for developing downstream AI based analysis of microscopy images.

This paper presents a novel approach for auto focusing a digital microscope, which utilises the recent advances in convolutional neural networks [10]. It can be implemented on any standard digital microscope with automated 3-axis control of the microscope stage. The method was trained on and results demonstrated using an open data set [9]. It is shown to achieve superior accuracy of focus distance estimation compared to the existing results [9] on the same data set across multiple types of staining protocols. It is also shown to be less sensitive to sample type and stain quality variation. Further, the method is shown to achieve superior performance with incoherent white light images, and does not require coherent or narrow band illumination. Thus, it can be potentially implemented on cost effective digital microscopes.

The paper is organised as follows: Section 2 analyses related work in the field of digital microscopy. Section 3 presents the details of the proposed method. Experimental results are presented in Section 4. Finally, Section 5 concludes the paper.

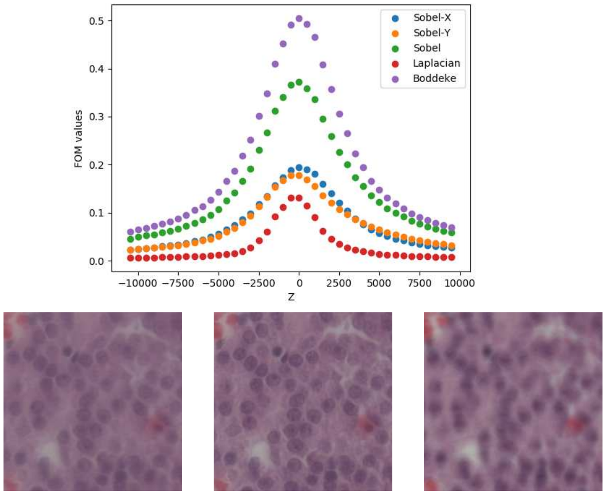
2. Related work

A typical WSI system uses a 20X objective lens (0.75NA) to capture images of the sample [1]. These images are aligned as tiles and stitched to produce the virtual slide image. The depth of field of such lenses are usually less than $1\mu m$. Both the topography of the biological sample, and the glass slide underneath can have depth variations. Thus, the microscope needs to be continuously focused as it moves from one field of view to another.

Auto focusing systems can be broadly divided into two categories – reflection based and image based [20]. In the reflection based method, an additional light source (often a laser diode) is introduced. The reflection of this source from the sample or the glass slide is used to estimate the focus distance. This method can perform rapid auto focusing. However, often a single reference depth is not sufficient to bring the entire field of view into focus. Presence of multiple reflections can also cause confusion.

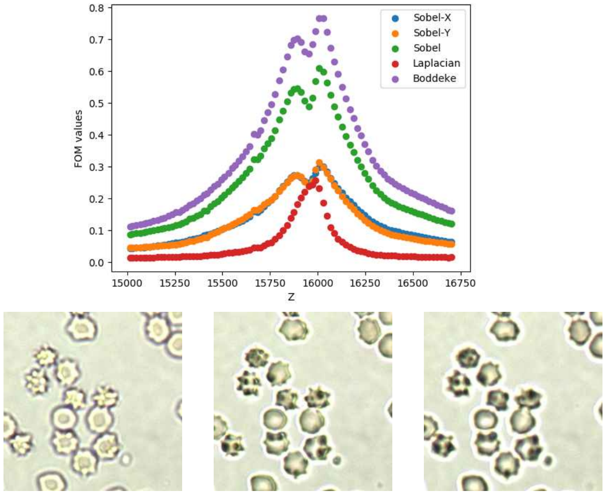
The image based auto focusing method involves capturing an image of the sample with a camera, through the objective lens, and then calculating a figure of merit (FOM) to judge the quality of focus. Multiple images are captured along the vertical axis, and the image with the best FOM value is taken as the "in focus" image.

Different FOM measures have been used in the literature, starting with [3]. Commonly used ones include norm of Sobel operator, variance of Laplacian, norm of Boddeke's operator, etc. A detailed description of these commonly used figures of merit can be found in [15, 4]. Figure 1 shows different figures of merit plotted across the focus stack for a single field of view in the data set provided by [9], along with a sample of images from the focus stack. It is evident from the figure that most FOMs peak at approximately the same position in the focus stack. However, the shape of the curve varies from one field of view to another. Primarily, it depends on the opacity of the objects, the number of objects, background illumination intensity, etc. Thus, figures of merit obtained from different fields of view are not comparable. The other disadvantage of FOM curves is that they can sometimes show local maxima, or multiple false peaks. This is especially true for unstained biological samples, which are more transparent in nature. See Figure 2 for an example of such false peaks. The image stack is captured on from a wet mount slide containing whole blood sample diluted with isotonic saline solution (0.85% NaCl). In this case, all the FOMs, except for variance of Laplacian, show two peaks, both of which are false (in the sense that the image does not visually appear "well focused" at these positions). The images corresponding to the two peaks are shown in Figure 2. Both false peaks occur due to the presence of high contrast optical artefacts around boundaries of cells. In some other cases, the variance of Laplacian FOM is also observed to exhibit multiple peaks across the focus stack, and often the highest peak is not the true one.



$D = -4.5\mu m$ $D = -2.5\mu m$ $D = 4.5\mu m$

Figure 1. The figure on the top plots FOM values for different distances, for a field of view taken from data set [9]. Sample images from the focus stack are shown, with the distance to the optimal focus position denoted as D.



Lower peak True focus Upper peak

Figure 2. A wet mount slide of blood mixed with saline. The FOM curves show false peaks.

Image based auto focusing systems typically employ one or more FOM, and tries to find the peak of the FOM. This is achieved through several methods:

- Uniformly sampling across the focus stack with fixed step sizes and then selecting the image with the peak FOM as the "in focus" image. This is a time consuming process.

- Using an adaptive step size while moving. A larger step size is taken when the FOM has a low magnitude. Step size is reduced as the system moves closer to the peak. Though faster than the uniform sampling approach, this method also requires multiple images to be captured. It also requires back tracking in case the selected step size causes the system to "jump over" the FOM peak position. The mechanical backlash [2], present in all microscope stages, complicates the backtracking process as positions are not exactly reproducible.

- A curve fitting based approach as in [20]. The shape of the FOM curve is estimated based on only a small set of samples. The system is moved directly to the estimated peak of the FOM curve. [20] shows the robustness of the proposed algorithm for various types of biological samples.

Liao *et al*. [11] use a novel approach to estimate the focus plane of a whole slide in an efficient manner. They use two green LEDs at a known spatial distance for illumination. When the sample is at a suboptimal focus distance and is illuminated with the two LEDs, the camera essentially captures two spatially separated copies of the image. The autocorrelation distance between the two copies is directly related to the defocus distance, and thus the defocus distance can be estimated without moving the sample vertically. They show good results on various sample types. While this approach is fast and accurate, it requires additional hardware in the form of extra LEDs which need to be aligned carefully.

In the recent past, convolutional neural networks (CNNs)[10] have been shown to be effective for several types of computer vision applications. This includes object recognition (e.g. [19, 18]), object localisation [14] and segmentation [7]. They have been successfully applied for classification [5, 13] and segmentation [17] of microscopic images as well.

Jiang *et al*. [9] explore the use of CNNs for the purpose of microscope auto focusing. The expectation is that given a sample image anywhere in the focus stack, the trained CNN should be able to estimate the optimal distance to be moved vertically (either up or down) to reach the optimal focus position. They train a deep CNN as a regression network. The training and test data set are publicly available. The test data consists of two different types of samples, prepared with different protocols at different sites. This tests the generalisation ability of the trained models in the face of sample type and stain colour variation.

It was observed in [9] that the trained CNN has relatively poor performance on the test set, when trained with only the RGB images. This is more evident on the test set with different protocol of preparation. To counter this, they propose the usage of *spectral domain* and *multi domain inputs*. In one of the experiments, the green channel of the captured image, along with its Fourier spectrum and phase, are used as the input to the network. The underlying idea is that the Fourier spectrum captures the frequency content of an image, and the frequency content is inversely related to the focusing distance. This approach is shown to produce superior results.

The use of the Fourier domain in CNNs have been explored earlier [12, 16]. Convolution in the spatial domain is equivalent to pixel-wise multiplication in the Fourier domain. Most of these approaches primarily use the Fourier domain to speed up the computationally expensive convolution step and replace it with the cheaper multiplication step. However, the absence of a meaningful non-linearity or activation function in the spectral domain necessitates conversion back into the spatial domain after each convolution layer [16]. The cost of repeated conversions between the Fourier and spatial domains partially negates the performance gain achieved by replacing the convolution step.

In [9], on the other hand, spatial convolution is used on the Fourier domain signal, which is theoretically unsound, even though practical results may be promising. Pixels in the Fourier spectrum have positional significance. CNNs, with shared weights in the convolution layers, are good at learning spatial patterns irrespective of their position. The same pattern appearing in different positions in the Fourier spectrum have different meanings.

3. Methods

This paper presents a deep regression convolutional neural network structure for auto focusing using spatial domain only inputs. Results are shown to be superior to those reported in [9] on the public data set.

3.1. Data set

The data set [9] consists of three different types of images:

- Images captured with a white light emitting diode (LED) illumination (termed as incoherent RGB input).

- Images captured with two green LEDs with a known spatial separation.

- Images captured with a single green LED.

In addition, the multi domain inputs for the above images are also provided. For each field of view (FOV), approximately 40 images are captured with defocus distance varying from $-10\mu m$ to $+10\mu m$ in steps of $0.5\mu m$. A total of

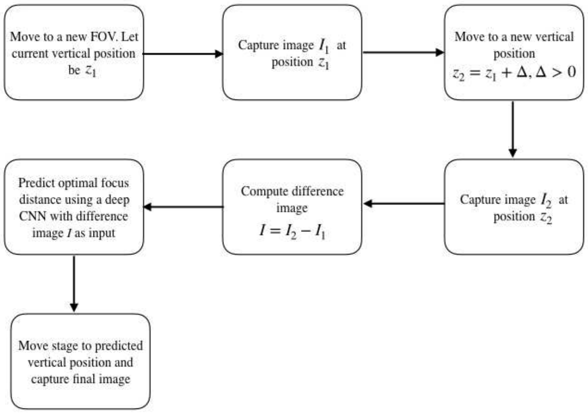

Move to a new FOV. Let current vertical position be z_1
Capture image I_1 at position z_1
Move to a new vertical position $z_2 = z_1 + \Delta, \Delta > 0$
Capture image I_2 at position z_2
Compute difference image $I = I_2 - I_1$
Predict optimal focus distance using a deep CNN with difference image I as input
Move stage to predicted vertical position and capture final image

Figure 3. Pictorial representation of the proposed method

128,699 patches of size 224 × 224 are available for training. For this work, the data is split into a 102,960 training set and a 25,739 validation set, ensuring that the two sets contain images from *different* samples.

The test set consists of two types of samples – one prepared with the same staining protocol as the training set, and the other prepared using a different staining protocol at a different site. As in [9], the test images are split into smaller patches of 224 × 224. The median of the predicted focusing distance across all patches from an image is taken as the prediction for that image. The median helps in avoiding outliers (which primarily come from the empty regions of an FOV, or regions with little or no contrast).

3.2. Proposed method

Jiang *et al*. [9] uses a single image, from anywhere in the focus stack, to estimate the distance to the optimal focus position, using a deep convolutional neural network. It was shown that using only a spatial domain RGB image yields suboptimal results when tested with a new test set having different staining properties.

This work proposes a method which uses the *difference* between two images in the focus stack to predict the distance to the optimal focus position. The flow of the proposed method is shown in Figure 3. The steps of the method can be described as follows:

1. At every new field of view, the process starts with a known vertical position. Let us call this position z_1. An image I_1 is captured at this position z_1. Note that the distance to the optimal focus position is not known at this time.

2. The stage is then moved *upwards* by a known distance Δ to a new position $z_2 = z_1 + \Delta$. Again, an image (I_2) is captured at position z_2.

3. The distance Δ should be greater than the depth-of-field of the microscope, yet not large enough to cause a significant difference in image properties between the two positions.

4. A *difference image* I is computed as $I = I_2 - I_1$.

5. The difference image I is fed as input to a deep CNN to predict the optimal focus position. I is pre-processed as described in Section 3.4 before being fed to the model.

The above method is easy to implement in a practical system. It has the overhead of one extra physical movement (by Δ) and one extra image capture. However, the improvement in prediction accuracy outweighs the overhead. The advantages of using this difference image I to predict the distance to the optimal focus position are described next.

3.3. Properties of the difference image

Using the difference of two images has several advantages:

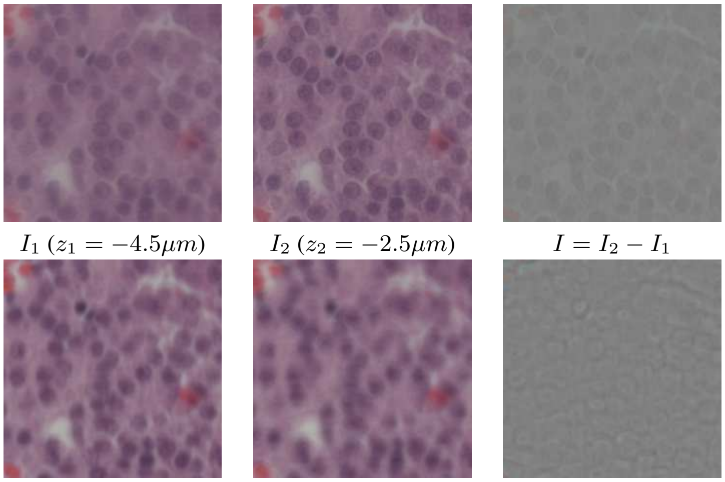
- It eliminates, to a large extent, the coarse colour information in the image, which is known to be a major source of over fitting [13].

- It emphasises local variations, which are important to gauge the defocus distance, and suppresses global features.

- The *defocus direction* is encoded in the difference image, if z_1 and z_2 are on the same side of the optimal focus position. To see how this happens, suppose both z_1 and z_2 are below the optimal focus position. Since z_1 is at a greater distance from optimal focus, the object features in I_1 will be more spatially dispersed. On the other hand, if both are above the optimal focus position, then I_2 will be more defocused and thus have more dispersed spatial features. The distinction between the two scenarios above will be evident in the difference image as the object edges will have different signs in the two cases.

- If z_1 and Δ are chosen such that z_1 and z_2 are always below the focus plane of the microscope, this eliminates backlash, as the stage does not need to change direction to get to the optimal position from z_2. Alternatively, if z_2 is predicted to be above the focal plane, the system could take a large step downwards, so as to bring it below the focal plane. Recomputing I_1 and I_2 from this new position will also eliminate backlash.

The primary disadvantage of the difference step is that it also amplifies the local noise in an image. The overall dynamic range of the image is reduced significantly due to the difference operation between two similar images. However, the random noise amplitude is not necessarily reduced. The noise thus becomes more prominent as compared to the single image input scenario.


I_1 ($z_1 = -4.5\mu m$) I_2 ($z_2 = -2.5\mu m$) $I = I_2 - I_1$

I_1 ($z_1 = 2.5\mu m$) I_2 ($z_2 = 4.5\mu m$) $I = I_2 - I_1$

Figure 4. Two examples of the difference image. The top row shows two images with defocus -4.5 and $-2.5\mu m$, and the difference of the two. Similarly, the second row shows images with defocus 2.5 and $4.5\mu m$, and their difference. Notice the distinct difference in the properties of the edges between the two difference images.

3.4. Image pre-processing

The image I is pre-processed as follows before being fed to the CNN:

1. The images I_1 and I_2 are smoothed with a median filter of size 3×3 to reduce local noise prior to the difference operation. Multiple filter sizes were tried. The 3×3 size produced best results.

2. The difference image I is again smoothed with the same filter.

3. A channel-wise local contrast normalisation is done on the smoothed difference image. Each channel in the image is centred to 0 (by subtracting the mean of the channel), and divided by the corresponding standard deviation. It is observed that the two test sets differed significantly – the variances of the 3 channels is very different in the two sets. The normalisation step above helps counter this difference. Our observation is that it also helps in better prediction on images where the dynamic range is very less – for example on smooth regions of the tissue.

3.5. Model architecture

A deep CNN based on the MobileNetV2 [18] architecture is used in this work. Recently, many "light weight" CNN architectures have been proposed [18, 8]. They have low computational cost and memory footprint. This makes them suitable for inferencing tasks on edge devices with low compute power. The proposed auto focusing system is likely to be deployed on an edge device (an automated microscope). Hence the choice of the base architecture. This work uses the MobileNetV2 network, except the topmost classification layer.

The model is a linear regression network. The output of the last feature map of MobileNetV2 is flattened and fed into a dense layer with a single output and no activation function. This model is trained with the mean squared error loss. An $L2$ regularisation term is added to the cost function for the last regression layer.

For the purpose of this work using the above data set, Δ was chosen as $2\mu m$. $2\mu m$ is greater than the depth-of-field in the microscope used for generating the data set (approximately $1\mu m$) and yet it is not big enough to cause significant change in image properties. For an image I_1 in the training set, with defocus distance z_1, we choose I_2 as the image at defocus distance $z_1 + 2\mu m$ (if it exists in the data set). The image I is then computed as $I = I_2 - I_1$ and the *corresponding defocus distance* is taken as z_2. If I_2 does not exist in the data set, I_1 is not used for training. The same convention is used for test images as well.

4. Experimental results

The results on the incoherent white light images from the data set [9] are presented.

4.1. Model training

Two versions of the base MobileNetV2 network were used – a pre-trained network (with weights initialised by training on the ImageNet [6] dataset), and a "fresh" network with the same architecture but randomly initialised weights. All models were trained with stochastic gradient descent, with learning rate 0.001. No learning rate annealing was used. Both versions of the base network were observed to have similar training time and performance. All results are reported on the “fresh” version of the network. The networks with difference image as input trained 2X faster and had significantly lower validation error than the ones trained on the raw RGB images.

4.2. Infrastructure

A machine with 6-core Intel Xeon 2.6GHz processor, 60GB RAM, and a single NVIDIA Tesla K80 GPU with 12GB memory was used for this work. Software included the Linux operating system (Ubuntu 16.04), NVIDIA Cuda 9.0, CUDNN 7.7. The Keras deep learning package (version 2.2.4) was used with Tensorflow (version 1.11.0) backend.

4.3. Single image input

For fair comparison with [9], in addition to the proposed difference method, the results of the regression network trained with a single input are presented in Table 1. The images were preprocessed with a channel-wise normalisation operation as described in Section 3.4, before being used for training or test. The results are expressed in terms of focusing error, i.e. the absolute difference between the predicted focusing distance and the ground truth. The test images are larger in size (1224×1024). They are split into tiles of size 224×224 pixels. The *median* of the predicted focus distance for these tiles is taken as the focusing distance of the overall image.

The network performs significantly better on the RGB only versions compared to [9]. The conjecture is that this is attributable to the greater depth of the network used and the normalisation operation. However, the results fall short of the best overall figures reported in [9] on the test set prepared with different protocol. The image-wise comparison is presented in Tables 3 and 4.

4.4. Difference image input

The results of the network trained with the difference image input are presented in Table 2. The current approach is found to outperform the best overall performance reported in [9]. The performance on the test images prepared with the same protocol is nearly the same as the network trained with single input only. However, there is a significant improvement in the performance on the test images prepared with different protocol. This shows the ability of the difference image to suppress non-essential features of the image while learning and thereby improving generalisation ability.

For the test images prepared with same protocol, the 98^{th} percentile error is $0.97\mu m$. The maximum error is $1.23\mu m$. For the test images prepared with different protocol, the 90^{th} percentile error is $0.96\mu m$ whereas the 95^{th} percentile error is $1.16\mu m$. The maximum error is $2.98\mu m$.

The image-wise comparison is presented in Tables 5 and 6.

4.5. Model optimisation

Since the target application for this system is an edge device, attempt is made to reduce the resource requirement of the model without affecting its performance to a great extent. The α (width multiplier) parameter of MobileNetV2 is fine tuned to arrive at an optimal balance between prediction accuracy and model size. In addition, the input image size is also varied.

A grid search is conducted over $\alpha = (0.25, 0.4, 0.5, 0.6, 0.75)$ to find the smallest α which gives a prediction accuracy on the test set prepared with different protocol, which is no worse than the best overall results in [9]. For $\alpha = 0.4$ the prediction accuracy is $0.31 \pm 0.24\mu m$ for the test set with same protocol, and $0.5 \pm 0.44\mu m$ for the test set with different protocol. Both values are below the ones in [9]. For smaller α (0.25), the performance on the second test set $(0.61 \pm 0.53\mu m)$ is slightly worse than the set limit. Higher values of α yielded little incremental benefit. The model size with $\alpha = 0.4$ has approximately $500,000$ parameters. This is a 4X decrease from the full MobileNetV2 model with the regression layer, which has approximately 2.2 million parameters.

To further reduce the compute requirements of the model, three different input sizes were also used – 56×56, 112×112 and 168×168 – and a model trained for each input size, with $\alpha = 0.4$. The selection criterion for the optimal size is the same as above. It is observed that the model with input size 56×56 is significantly faster, and its performance on the first test set is acceptable. However, the performance $(0.65 \pm 0.81\mu m)$ on the second test set fell below the limit. The model trained with input size 112×112 met the performance criteria. Use of 112 input size causes a 4X decrease in feature map size over the original 224×224 input. The execution performance of the original model compared with the model with $\alpha = 0.4$ and input size 112×112 is shown in Table 7. It can be seen that there is a significant reduction in GPU memory usage and execution time. Input size of 168×168 didn’t yield any incremental benefit.

5. Conclusion and future work

This paper presented a new method for applying deep learning for focus distance estimation in automated digital microscopy. The method is shown to be superior to results in existing literature in terms of generalisation error over multiple staining protocols. The method is also easy to practically implement on any automated microscope. The compute resource requirement of the proposed model is shown to be low, making it possible to be deployed on an edge device with limited resources.

Test set	Current work (RGB only) (μm)	Jiang *et al.* [9] (Single domain) (μm)	Jiang *et al.* [9] (best overall) (μm)
Same protocol	$\mathbf{0.25 \pm 0.23}$	0.50 ± 0.32	0.46 ± 0.34
Different protocol	0.62 ± 0.79	1.94 ± 1.91	$\mathbf{0.53 \pm 0.59}$

Table 1. Focusing errors (absolute difference between predicted and ground truth focusing distance for an image) obtained using RGB only images for training, compared with those presented in [9]. The best overall figures for [9] refer to results on incoherent illumination images only, whether RGB only or multi domain. Figures represented as mean error ± standard deviation of error.

Test set	Current work (difference image) (μm)	Jiang *et al.* [9] (Single domain) (μm)	Jiang *et al.* [9] (best overall) (μm)
Same protocol	$\mathbf{0.22 \pm 0.25}$	0.50 ± 0.32	0.46 ± 0.34
Different protocol	$\mathbf{0.36 \pm 0.37}$	1.94 ± 1.91	0.53 ± 0.59

Table 2. Focusing errors (absolute difference between predicted and ground truth focusing distance for an image) obtained using difference images images for training, compared with those presented in [9]. The best overall figures for [9] refer to results on incoherent illumination images only, whether RGB only or multi domain. Figures represented as mean error ± standard deviation of error.

Tissue sample	Num. images	Current work (RGB only) (μm)	Jiang *et al.* [9] (Single domain) (μm)
Sample 1	164	$\mathbf{0.25 \pm 0.23}$	0.33 ± 0.25
Sample 2	82	$\mathbf{0.33 \pm 0.20}$	0.33 ± 0.26
Sample 3	41	$\mathbf{0.26 \pm 0.14}$	0.37 ± 0.22
Sample 4	246	$\mathbf{0.21 \pm 0.19}$	0.53 ± 0.28
Sample 5	82	$\mathbf{0.25 \pm 0.28}$	0.58 ± 0.31
Sample 6	82	$\mathbf{0.38 \pm 0.31}$	0.87 ± 0.57

Table 3. Comparison of focusing error for each individual sample in the test set prepared with the *same protocol*. As in Table 1, the figures are expressed as $\mu \pm \sigma$ of the focusing error.

Tissue sample	Num. images	Current work (RGB only) (μm)	Jiang *et al.* [9] (Single domain) (μm)
Sample 7	41	0.63 ± 0.44	$\mathbf{0.48 \pm 0.32}$
Sample 8	123	$\mathbf{0.65 \pm 1.80}$	1.32 ± 1.29
Sample 9	205	$\mathbf{0.46 \pm 0.40}$	2.69 ± 2.41
Sample 10	246	$\mathbf{0.98 \pm 0.75}$	2.19 ± 2.15
Sample 11	246	$\mathbf{0.46 \pm 0.33}$	2.83 ± 3.25
Sample 12	205	$\mathbf{0.72 \pm 0.68}$	1.00 ± 0.77
Sample 13	246	$\mathbf{0.55 \pm 0.48}$	2.02 ± 2.48

Table 4. Comparison of focusing error for each individual sample in the test set prepared with the *different protocol*. As in Table 1, the figures are expressed as $\mu \pm \sigma$ of the focusing error.

Tissue sample	Num. images	Current work (μm)	Jiang *et al.* [9] (best overall) (μm)
Sample 1	164	$\mathbf{0.20 \pm 0.22}$	0.27 ± 0.18
Sample 2	82	$\mathbf{0.21 \pm 0.30}$	0.70 ± 0.83
Sample 3	41	$\mathbf{0.28 \pm 0.23}$	0.31 ± 0.22
Sample 4	246	$\mathbf{0.19 \pm 0.19}$	0.42 ± 0.24
Sample 5	82	$\mathbf{0.29 \pm 0.25}$	0.36 ± 0.29
Sample 6	82	0.42 ± 0.32	$\mathbf{0.45 \pm 0.24}$

Table 5. Comparison of focusing error for each individual sample in the test set prepared with the *same protocol*. The figures for [9] are the best overall across all approaches explored for incoherent illumination images. Thus, all figures do not refer to the same methodology. As in Table 2, the figures are expressed as $\mu \pm \sigma$ of the focusing error.

Tissue sample	Num. images	Current work (μm)	Jiang *et al.* [9] (best overall) (μm)
Sample 7	41	$\mathbf{0.28 \pm 0.24}$	0.48 ± 0.32
Sample 8	123	$\mathbf{0.68 \pm 0.38}$	0.99 ± 1.51
Sample 9	205	$\mathbf{0.25 \pm 0.27}$	0.28 ± 0.28
Sample 10	246	$\mathbf{0.35 \pm 0.48}$	0.38 ± 0.38
Sample 11	246	$\mathbf{0.37 \pm 0.29}$	0.43 ± 0.69
Sample 12	205	0.55 ± 0.39	$\mathbf{0.52 \pm 0.29}$
Sample 13	246	0.33 ± 0.33	$\mathbf{0.29 \pm 0.22}$

Table 6. Comparison of focusing error for each individual sample in the test set prepared with the *different protocol*. The figures for [9] are the best overall across all approaches explored for incoherent illumination images. Thus, all figures do not refer to the same methodology. As in Table 1, the figures are expressed as $\mu \pm \sigma$ of the focusing error.

Future work will involve extending this method to other types of biological samples, especially to liquid samples, which have multiple sharpness peaks and focusing is known to be more challenging.

One of the observed drawbacks of the proposed method is its sensitivity to local noise. It was noticed that compression artefacts have a significant influence on the prediction accuracy. A lower quality of JPEG compression causes the

Model	# Params	CPU time (seconds)	GPU memory (GB)	Focusing error Set 1 (μm)	Focusing error Set 2 (μm)
Original model	2,200,000	270	2.17	0.22 ± 0.25	0.36 ± 0.37
Reduced model	500,000	100	0.98	0.32 ± 0.27	0.43 ± 0.36

Table 7. Comparison of performance of the original model with the reduced version ($\alpha = 0.4$ and input size 112×112). The execution time is measured over model evaluation for 26,240 patches from the second test set. Batch size used is 8. The time reported includes the time required for reading the image from disk, resizing, and splitting the larger images into patches on the fly. Details of the hardware and software specs are mentioned in Section 4.1.

prediction accuracy to deteriorate drastically. On the other hand, the models trained with single image input were observed to be more robust to JPEG compression artefacts. This issue will also be addressed in future.

Acknowledgement

The author would like to thank Shaowei Jiang for helping out with numerous questions related to the data set and also providing the uncompressed version of the images. Thanks are also due to Debdoot Sheet for his help in reviewing the draft.

References

[1] E. Abels and L. Pantanowitz. Current state of the regulatory trajectory for whole slide imaging devices in the usa. *Journal of pathology informatics*, 8, 2017. 1, 2

[2] F. R. Boddeke, L. J. Van Villet, and I. T. Young. Calibration of the automated z-axis of a microscope using focus functions. *Journal of microscopy*, 186(3):270–274, 1997. 3

[3] 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. *Journal of Histochemistry & Cytochemistry*, 24(1):100–111, 1976. 1, 2

[4] M. A. Bueno-Ibarra, J. Alvarez-Borrego, L. Acho, and M. C. Chavez-Sanchez. Fast autofocus algorithm for automated microscopes. *Optical Engineering*, 44(6):063601, 2005. 2

[5] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber. Mitosis detection in breast cancer histology images with deep neural networks. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 411–418. Springer, 2013. 1, 3

[6] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. *Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 5

[7] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. *IEEE conference on computer vision (ICCV)*, pages 2980–2988, 2017. 3

[8] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and $<$ 0.5 mb model size. 2016. arXiv:1602.07360. 5

[9] 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. *Biomedical optics express*, 9(4):1601–1612, 2018. 1, 2, 3, 4, 5, 6, 7

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing system*, pages 1097–1105, 2012. 1, 3

[11] J. Liao, Y. Jiang, Z. Bian, B. Mahrou, A. Nambiar, A. W. Magsam, K. Guo, S. Wang, Y. ku Cho, and G. Zheng. Rapid focus map surveying for whole slide imaging with continuous sample motion. *Optics letters*, 42(17):3379–3382, 2017. 3

[12] M. Mathieu, M. Henaff, and Y. LeCun. Fast training of convolutional networks through ffts. 2013. arXiv:1312.5851. 3

[13] D. Mundhra, B. Cheluvaraju, J. Rampure, and T. R. Dastidar. Analyzing microscopic images of peripheral blood smear using deep learning. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 178–185. Springer, 2017. 1, 3, 4

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. *IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 3

[15] R. Redondo, G. Cristóbal, G. B. Garcia, O. Deniz, J. Salido, M. del Milagro Fernandez, J. Vidal, J. C. Valdiviezo, R. Nava, B. Escalante-Ramírez, et al. Autofocus evaluation for brightfield microscopy pathology. *Journal of biomedical optics*, 17(3):036008, 2012. 2

[16] O. Rippel, J. Snoek, and R. P. Adams. Spectral representations for convolutional neural networks. *Advances in neural information processing systems*, pages 2449–2457, 2015. 3

[17] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241, 2017. 3

[18] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. *IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 3, 5

[19] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. 2014. arXiv:1409.1556. 3

[20] 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. *Optics express*, 16(12):8670–8677, 2008. 1, 2, 3