DEEP AUTOFOCUSING FOR DIGITAL PATHOLOGY WHOLE SLIDE IMAGING

DEEP AUTOFOCUSING FOR DIGITAL PATHOLOGY WHOLE SLIDE IMAGING

BY

QIANG LI, M.S.

A THESIS

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL & COMPUTER ENGINEERING

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

McMaster University © Copyright by Qiang Li, April 2024

McMaster University

DOCTOR OF PHILOSOPHY (2024)

Hamilton, Ontario, Canada (Electrical and Computer Engineering)

TITLE:	DEEP AUTOFOCUSING FOR DIGITAL PATHOLOGY WHOLE SLIDE IMAGING
AUTHOR:	Qiang Li M.S. (Instrument Science and Technology), Harbin Institute of Technology, Harbin, China
SUPERVISOR:	Dr. Xiaolin Wu
NUMBER OF PAGES:	xix, 150

Abstract

The quality of clinical pathology is a critical index for evaluating a nation's healthcare level. Recently developed digital pathology techniques have the capability to transform pathological slides into digital whole slide images (WSI). This transformation facilitates data storage, online transmission, real-time viewing, and remote consultations, significantly elevating clinical diagnosis. The effectiveness and efficiency of digital pathology imaging often hinge on the precision and speed of autofocusing. However, achieving autofocusing of pathological images presents challenges under constraints including uneven focus distribution and limited Depth of Field (DoF). Current autofocusing methods, such as those relying on image stacks, need to use more time and resources for capturing and processing images. Moreover, autofocusing based on reflective hardware systems, despite its efficiency, incurs significant hardware costs and suffers from a lack of system compatibility. Finally, machine learning-based autofocusing can circumvent repetitive mechanical movements and camera shots. However, a simplistic end-to-end implementation that does not account for the imaging process falls short of delivering satisfactory focus prediction and in-focus image restoration.

In this thesis, we present three distinct autofocusing techniques for defocus pathology images: (1) Aberration-aware Focal Distance Prediction leverages the asymmetric

effects of optical aberrations, making it ideal for focus prediction within focus map scenarios; (2) Dual-shot Deep Autofocusing with a Fixed Offset Prior is designed to merge two images taken at different defocus distances with fixed positions, ensuring heightened accuracy in in-focus image restoration for fast offline situations; (3) Semi-blind Deep Restoration of Defocus Images utilizes multi-task joint prediction guided by PSF, enabling high-efficiency, single-pass scanning for offline in-focus image restoration.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor Dr. Xiaolin Wu for his continuous support throughout my Ph.D study, for his patience, enthusiasm, and immense knowledge. Without his invaluable guidance and dedication, none of the thesis work would have materialized. I could not have imagined having a better supervisor for my Ph.D study.

I would like to thank my committee members Dr. Jun Chen and Dr. Xun Li for giving me guidance and support over the years. I would also like to thank Dr. Junsong Yuan for being my external examiner.

My gratitude also goes to my friends and fellow collaborators, especially to Xianming, Yanhui, Fangzhou, Hao Xu, Hong Chen, Liangyan, Sitong and Allen, for generously sharing their advice and knowledge. I appreciate the friendship built among us and I wholeheartedly believe that our friendships will last even beyond this journey.

Last but not least, I would like to thank my family for their unwavering support throughout my PhD journey. To them, I dedicate this thesis.

Table of Contents

List of Figures

List of Tables

Abbreviations

WSI	Whole Slide Image/Imaging
AI	Artificial intelligence
ML	Machine Learning
CNN	Convolutional Neural Network
DL	Deep Learning
DNN	Deep Neural Network
GT	Ground Truth
GAN	Generative Adversarial Network
GPU	Graphics Processing Units
MSE	Mean Squared Error
MAE	Mean Absolute Error
SD	Standard Deviation
PSNR	Peak Signal-to-noise Ratio

SSIM	Structural Similarity Index
SOTA	State of the Art
PSF	Point Spread Function
OTF	Optical Transfer Function
DoF	Depth of Field
NA	Numerical Aperture
FoV	Field of View

Chapter 1

Introduction

1.1 Whole Slide Imaging

Whole slide imaging (WSI), also referred to as *virtual microscopy* [61, 84], is developed to transform the conventional microscope glass slides to splicing seamless digital images that can be analyzed on a computer, easily stored, and quickly shared with other researchers no matter where they are [25, 2]. In the medical realm, WSI continues to gain traction worldwide as a feasible approach for digital pathology. It has become a vital means gradually in biomedical research, clinical diagnosis and prognosis of diseases like cancer [19]. A remarkable milestone is that in 2017 the US Food and Drug Administration has approved Philips' WSI system for the primary diagnostic use [1].

A typical WSI process includes: 1) utilizing a scanner to digitize tiles of a sample, which generates digital images that are then stitched together to produce a complete and seamless representation of the original entire slide [91]; 2) employing specialized software to view and analyze these digital images [62]. It is clear that the quality

of captured images in the first step is critical for the performance of WSI system. A fundamental challenge in WSI is how to produce a high-quality, in-focus image at fast speed. Specifically, a whole slide scanner is essentially a microscope with a high-resolution objective lens (typically larger than 0.75 NA), whose DoF is usually less than 1 μm. The small DoF in WSI systems poses a challenge to acquiring in-focus images of tissue sections of uneven topography [36]. The out-of-focus blurring artifact is the main source of image quality degradation in WSI [34]. In addition, the use of high NA objectives results in a very small field of view (FoV), with each title being only 2500 square micrometers. A typical pathology specimen of $1.5 \times 1cm^2$ can consist of as many as 6000 such titles. Clearly, bringing all these tiles into focus one by one creates a severe bottleneck to the throughput of a WSI system.

The process referred to as autofocusing is conducted to solve this problem. In the literature, one popular solution for autofocusing is the so-called focus map surveying method [42]. It creates a focus map before scanning. More specifically, for each tile (a point in the focus map), a z-stack of images of different focal distances is taken. The sharpest image in the z-stack [90], identified by a contrast or entropy criterion, determines the focus point for the tile. This process is repeated for all tiles of the entire tissue slide to generate the focus map. According to this focus map, the mechanical system scans the sample and performs in-focus tile-by-tile shooting. However, there are two drawbacks to this focus map surveying method. First, as stated above, for each tile the system takes as many as N images, N being the depth of the z-stack, which is time-consuming. Creating a focus map for the slide is a significant overhead. While selecting a subset of tiles for focus point surveying can save time to some extent, it compromises the accuracy of focus. Second, the system

needs to make two passes of the slide. The first pass is to generate the focus map; the second pass is to shoot tiles one by one according to the focus map. Making an extra pass slows down the image acquisition as moving between the files incurs mechanical acceleration and deceleration. In order to achieve rapid autofocusing, some works consider using additional hardware. For instance, the dual-camera setup is proposed in [56], in which a secondary high-speed camera is employed to acquire images to avoid axial scanning. However, this approach is not feasible in the alignment of the additional camera to the microscope. Moreover, its compatibility with most existing WSI platforms remains open to question.

Considering the limitations of the conventional methods, some researchers have begun to investigate the possibility of exploiting advanced machine learning algorithms to solve the autofocusing problem. The work in [30] is the first one that uses deep convolution neural networks (CNNs) to predict the focal position, which acquired ∼130,000 images with different defocus distances as the training dataset, and used an end-to-end deep residual network to bulid the connection between the input image and its focal distance. This approach can capture images on the fly without focus map surveying. Despite this method achieves remarkable autofocusing performance, methodologically it is not easy to derive a model that accurately describes the relationship between an image with complex contents and a numerical value (the defocus distance). Pinkard *et al.* [64] also proposed to utilize CNNs to estimate focus distances, which emphasizes the lack of generalization across various sample types. Dastidar *et al.* [15, 66] explore the two-shot images as the input of CNNs for the purpose of focus distance estimation. Though this method improves the estimation accuracy, it needs the extra time to capture the second image. Wu *et al.* [87] proposed

a deep neural network to refocus a virtually two-dimensional fluorescence image onto user-defined three-dimensional (3D) surfaces within the sample. However, pathological images we work on are with more complex biological structures than fluorescent images. Thus, the autofocusing of pathological images is more challenging than that of fluorescent images. In summary, achieving focus prediction and in-focus restoration through a simplistic end-to-end approach without taking into account the principles of optical imaging proves to be challenging.

1.2 Contributions and Thesis Organization

To address the aforementioned limitations of existing autofocusing methods, this thesis proposes to incorporate advanced artificial intelligence (AI) techniques. Specifically, we overcome the following challenges:

- **Aberration-aware Focal Distance Prediction**: Employing machine learning techniques enables the prediction of focus distance, effectively circumventing the repetitive movements and focusing exposures inherent to traditional image stacking methods. However, optical aberrations inevitably cause pathological images with positive and negative defocus to exhibit distinct characteristic differences. Such inherent limitations of the imaging system reduce the accuracy of focus prediction. To address the aforementioned issue, this thesis proposes an aberration-aware focus prediction method by feature classification and focus regression. This method ingeniously leverages the characteristic differences caused by aberrations as a physical guide, specifically, positive defocus exhibiting striation artifacts and negative defocus featuring uniform blurring. The method

develops a binary classification network to differentiate samples with positive and negative focus shifts, leveraging the principle that defocus features in the same direction share similarities. Subsequently, utilizing the defocus data from both categories, it designs a regression network for focus prediction, forming a complete classification-regression deep cascade autofocusing network. Experimental evidence indicates that, relative to the baseline classification method without aberration guidance, our approach achieves a 26% reduction in focus prediction error. This method is suitable for scenarios such as constructing focus maps for focus prediction, where through distance prediction followed by system focusing and exposure, true in-focus images can be obtained.

- **Dual-shot Deep Autofocusing with a Fixed Offset Prior**: Directly using machine learning algorithms to deblur defocus pathological images is undeniably the most efficient approach, eliminating the need to capture dozens of images at different defocus distances as in traditional image stack methods. However, under constraints of high magnification objective lenses such as uneven focus distribution and limited DoF, blind deblurring pathological images is challenging. To tackle the aforementioned problem, this thesis proposes a dual-shot deep autofocusing with a fixed offset prior to achieve blind deblurring.

  This method designs an implicit position prior, utilizing two defocus images taken at fixed relative positions to derive a univariate equation for the in-focus image, thereby transforming the problem of blind deblurring into a non-blind deblurring issue. This approach uses only two images taken at different focal lengths but with relatively fixed positions. The dual-shot design helps to merge complementary information from both images, overcoming the challenges posed

by uneven focus distribution and DoF limitations in pathological samples. Experimental findings show that, compared to the baseline single-image method, our approach enhances image quality by 7%. This method is suitable for online scanning and offline recovery. With just two scanning exposures, it can produce high-quality in-focus images.

- **Semi-blind Deep Restoration of Defocus Images**: To speed up the dual-shot autofocusing method, we propose a one-shot autofocusing method. This one-shot method is similar to single-image deblurring, aiming to restore clear images through algorithmic reconstruction. Currently, existing single-image deblurring methods do not consider prior information related to the imaging system. Therefore, our method introduces a multi-task joint training strategy guided by the PSF prior, where the network simultaneously performs dual predictions to the in-focus image and the defocus image. We can regenerate re-defocus images by utilizing estimated blur kernel PSF. However, microscope PSFs are affected not only by defocus but also by aberrations (such as spherical aberration, chromatic aberration), and demosaicing effects. This complexity surpasses that of theoretical Bessel PSF functions, hence we utilize neural network prediction for the PSF mask rather than traditional optimization algorithms. To address color channel mismatches, we utilize Y channel data to predict the PSF mask. Subsequently, we achieve re-defocus images convoluted by the corresponding PSF from classification. Finally, the network can impose joint constraints on both in-focus and defocus images, thereby significantly enhancing image restoration performance. Experimental results demonstrate that, compared to the baseline method lacking PSF guidance, our approach results

in a 2.7% improvement in image quality. This method is suitable for scenarios involving online scanning and offline recovery. The one-shot scanning greatly improves scanning efficiency while meeting basic imaging quality requirements.

Chapter 2

Related Work

At present, there are numerous autofocusing methods available. This section provides an overview of representative autofocusing techniques for microscopic imaging along with corresponding examples. These are categorized as follows:

- **Hardware-based Reflective Autofocusing**: This includes methods like confocal pinhole detection, oblique illumination triangulation, and oblique illumination with weak coherent interference, among others.
- **Real-time Image-based Autofocusing**: Examples of this category encompass the z-stack autofocusing method, dual-sensor independent scanning, beam array technique, tilted sensor method, phase detection, and dual-LED illumination, to name a few.
- **Deep Learning-based Autofocusing**: This comprises methods such as focus prediction and focal plane recovery.

Finally, this section also touches upon the current state of research on WSI scanning strategies.

2.1 Hardware-based Reflective Autofocusing Techniques

Reflective-based autofocus aims to detect the axial position of a reference plane. Typically, this plane lies at the interface between glass and liquid, where cells often adhere, or at the air-glass boundary at the bottom of a cell culture container. During experiments, the focus drift correction system continually searches for the axial position of the reference plane. It maintains a consistent distance between the objective lens and the reference plane through an electric axial driver. This section delves into three reflective hardware-based autofocusing methods, namely: confocal pinhole detection, oblique illumination triangulation, and oblique illumination with weak coherent interference.

2.1.1 Confocal Pinhole detection

Liron *et al.* introduced a laser reflection autofocusing method using confocal pinhole detection [46]. The optical setup is depicted in Figure 2.1, where the expanded laser beam focuses on the substrate of the specimen (illustrated as the red beam). The light reflected from the substrate passes through the confocal pinhole to reach the photodetector (represented by the yellow beam). The fraction of laser intensity reflected at the interface roughly corresponds to the square of the refractive index difference. Reflections from the glass-air interface account for approximately 4% of the incident beam, while those from the glass-specimen interface are merely 0.4%. The inset of Figure 2.1 showcases intensity curves obtained by axially scanning the objective lens to various positions. The first pronounced peak correlates with the air-glass interface, and the

subsequent fainter peak pertains to the specimen-glass interface. Solid and dashed lines represent results for 100 μm and 200 μm pinholes, respectively. As indicated by the dashed line in Figure 2.1, enlarging the confocal pinhole size broadens the peak width. Such a modification reduces unwanted interference patterns, facilitating the data analysis process. The method employs a two-phase operation for autofocusing execution. The first phase, termed long peak detection search, involves high-speed axial scanning of the objective lens to identify the pronounced peak. Through the position of this first peak, the second peak's location can be estimated by factoring in the glass substrate's thickness. The second phase, dubbed local peak search, allows precise peak searching within a relatively shorter range.

While this confocal detection technique enables precise autofocusing, its primary drawback is the necessity of axial scanning to obtain the trajectory curve. Another limitation is the significant intensity disparity between the two peaks, with the weaker peak easily overshadowed by the first pronounced one, especially for objectives with lower magnification. In the second method of this section centered on reflective hardware techniques, we'll discuss a strategy to overcome the first method's shortcoming—locating the initial peak position without axial scanning. The third method explores another approach to address both drawbacks: reducing the signal strength of the first peak and pinpointing both peaks without axial scanning.

2.1.2 Oblique Illumination Triangulation

To locate the axial position of the interface without axial scanning, one can illuminate the specimen with tilted incident light and measure the lateral displacement of the reflected beam, as illustrated in Figure 2.2. The triangulation method for microscope

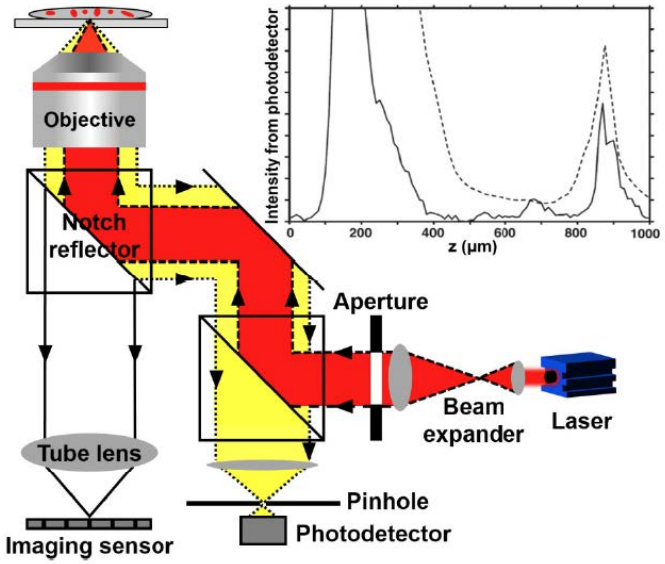

Figure 2.1: An autofocusing system using confocal pinhole detection [46]

autofocusing can be traced back to Reinheimer's 1973 patent, which proposed shaping the illumination beam to only occupy half the cross-section of the light pupil aperture [67]. When the specimen surface is positioned at different axial locations, the beam reflected from the surface will exhibit varying lateral displacements. Reflected light from the specimen surface is detected by two photodetectors for differential measurements. The differential signal detected by these two sensors drives the parfocal helix. For instance, if the specimen surface aligns with the focal plane, the reflected light is guided to the boundary of both photodetectors, producing a differential signal of zero, necessitating no adjustments. If the specimen surface is above the focal plane, the reflected light leans towards one of the photodetectors, and the resulting differential signal drives the specimen's moving platform. Conversely, if the specimen surface is below the focal plane, the differential signal from the two photodetectors propels the specimen stage upward. Similar schemes have also been suggested in

recent literatures [47, 48, 50, 49].

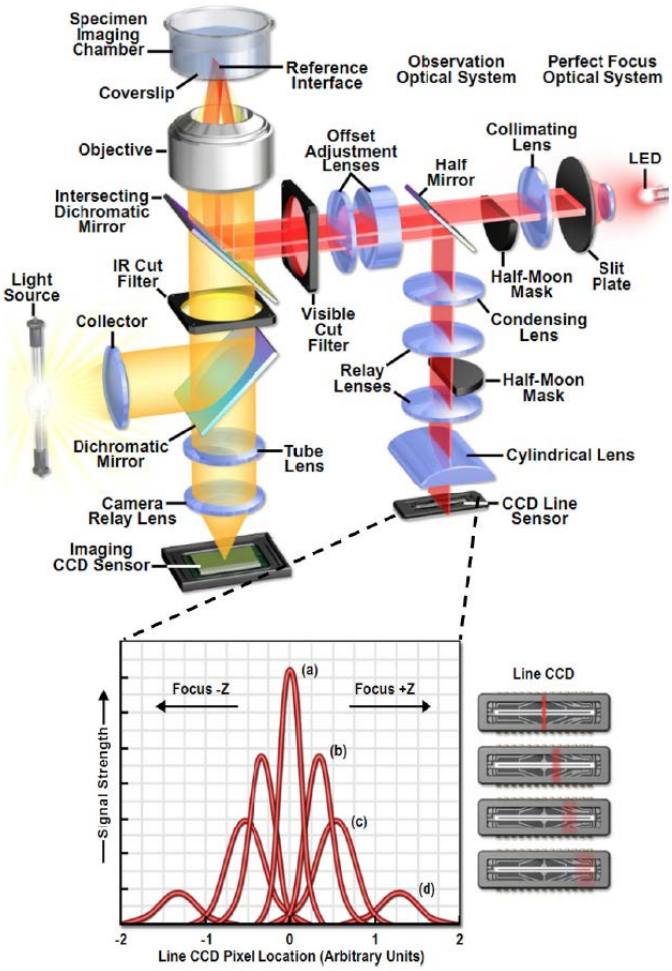

Figure 2.2: Nikon perfect focus system [67]

2.1.3 Oblique Illumination with Weak Coherent Interference

In a 1996 patent, Wei and Hellmuth introduced an autofocusing method using Optical Coherence Tomography (OCT), specifically utilizing an axial depth reflectivity device known as A-scan to determine the specimen position, as shown in Figure 2.3 [83]. In related patents, autofocusing for ophthalmic surgery microscopes was achieved

using a coaxial setup. However, this is not suitable for high-resolution imaging of pathological tissue slides covered with coverslips. A primary reason is the overlap between the strong reflection signal from the glass surface and the weak reflection signal from the specimen. Given the dominant reflection from the glass surface, positioning the specimen with sub-micron precision is challenging [10].

One solution to this problem is to significantly reduce the light reflected from the glass surface while maintaining the scattered light from the sample relatively constant. Figure 2.3 illustrates a solution employing an off-axis setup where light illuminates the sample at an inclined angle, ensuring light directly reflected from the glass surface won't couple back into the interference system. In Figure 2.3, a broadband superluminous diode is used as a low-coherence light source, with a spectrometer configured for Fourier domain OCT measuring axial depth reflectivity profiles. By Fourier-transforming the captured spectrum, the sample's position can be determined, by adjusting the objective lens to the focal point. Since OCT is highly sensitive to refractive index changes within the specimen, this method can handle transparent samples that might be challenging for traditional focus mapping techniques. Drawbacks include the complexity of the Fourier domain OCT setup, precise optical alignments, and the system's high maintenance requirements.

2.2 Real-time Image-based Autofocus Techniques

Before initiating system scanning, methods to create a focus map require obtaining a z-stack for the focus of each tile. This involves scanning the specimen to different x-y positions to acquire multiple z-stacks and subsequently generate the focus map. In most WSI sessions, the time spent creating this focus map constitutes a significant

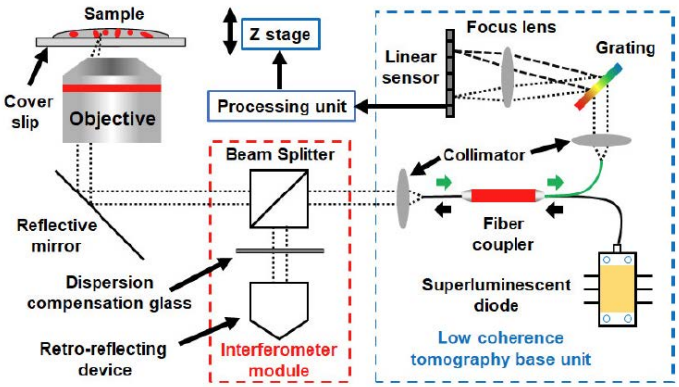

Figure 2.3: Low-coherence interferometry for reflective real-time autofocusing [83]

portion of the total scanning duration. In this section, we present six real-time autofocusing methods based on imaging. They include the z-stack autofocusing technique, dual-sensor independent scanning, split-beam array method, tilted sensor approach, phase detection method, and dual-LED illumination technique.

2.2.1 Z-stack Autofocusing Technique

The z-stack autofocusing method is depicted in Figure 2.4. This method revolves around acquiring a series of images before and after the focal plane, enabling the determination of the optimal focal length position by analyzing the focal positions of these images. When employing the z-stack technique for autofocusing, it's essential first to capture a series of images near the specimen's focal plane. These images cover a certain depth of focus by minutely adjusting the focal length. Subsequently, the optimal focal length can be determined by comparing the focal positions or image quality metrics of these images, resulting in a clear image. One of the method's strengths is its ability to cope with the unevenness and complexity of the sample surface, achieving precise autofocusing. By capturing a series of images and analyzing

their focal points, it eliminates the irregularities and uneven features on the sample surface, producing a universally clear image. In imaging systems like microscopes, the z-stack method is extensively employed to achieve high-quality autofocusing. It not only enhances image clarity and detail but also accelerates imaging speed and improves work efficiency. As such, the z-stack method has become one of the widely adopted autofocusing techniques in many labs and research fields. However, it also has some clear drawbacks:

(a) Time and Resource Consumption: Employing the z-stack method usually necessitates acquiring a series of images covering a certain depth of focus, which means spending more time and resources capturing and processing these images. Especially for larger or complex samples, a substantial number of images might be required to achieve optimal focus.

(b) Data Storage and Processing Demands: As the z-stack method involves capturing multiple images, it consumes more storage space. Moreover, processing and analyzing these images demand additional computational resources and algorithms to extract focal information. This could pose certain requirements for hardware and computational capacities, adding to the system's complexity.

(c) Motion Artifacts and Sample Movement: During z-stack autofocus, minute movements or vibrations might be present in the sample or the camera, potentially causing alignment issues between images. This could lead to motion artifacts or inaccurate focusing results. This issue might be even more pronounced for living samples or imaging processes that require a longer duration.

(d) Parameter Selection and Adjustment: The z-stack method necessitates the selection and tweaking of several parameters, such as step size, sampling intervals,

and focusing range. The choices made regarding these parameters can influence the focusing outcome, necessitating experimentation and optimization to determine the optimal parameter settings.

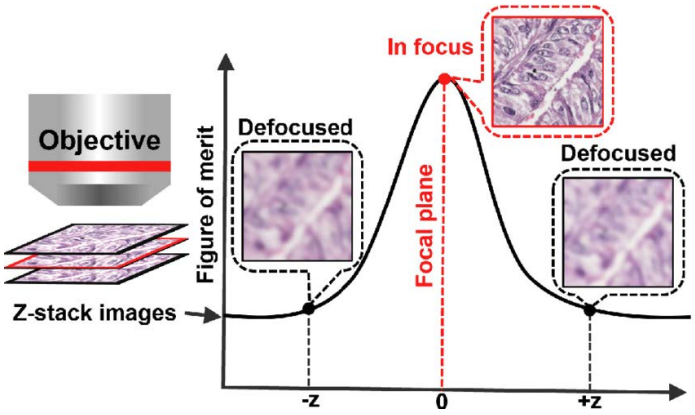

Figure 2.4: The traditional axial scanning z-stack procedure for autofocusing [78]

2.2.2 Dual-sensor Independent Scanning Technique

The traditional focal plane imaging method employs a single image sensor, both for measuring focus and capturing images. Between two successive image captures, there's a quantifiable "dead time" for reading data out to storage. As a result, during this "dead time," the camera cannot be used for focus measurements. Existing literature suggests the use of a separate auxiliary image sensor for parallel focus measurement [54].

Figure 2.5 illustrates the principle and operational process of the dual-sensor independent scanning concept [56]. As shown in Figure 2.5(a), the system employs an independent camera, termed the quasi-focal sensor, to measure focus, while the primary camera captures high-resolution images of pathological tissue samples. During the scanning process, the platform remains in constant motion, and short-pulse light

is utilized during imaging to eliminate motion blur. As depicted in Figure 2.5(b), the quasi-focal sensor captures three autofocusing images, each with a slightly different focal plane. Based on these three images, the system computes the optimal focal position and relocates the sample to this plane where the primary camera can capture a high-resolution image [90]. While the main camera is reading image data, the system repeats autofocusing for the next tile position, predicting its subsequent optimal focal plane. Since the platform remains in continuous motion throughout this process, the three captured focus images share only a small overlapping region, as demonstrated in Figure 2.5(c). Only this overlapping region can be used to compute the focal length. The autofocusing performance of the dual-sensor independent scanning system has been validated across various tissue sections. The continuous motion scheme averages a focus error of approximately $0.30\mu m$, with around 95% of local micro-images falling within the system's depth of field range.

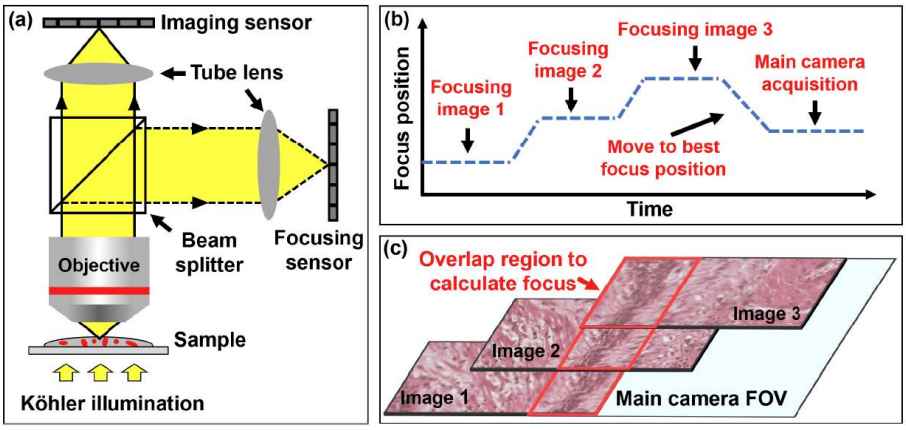

Figure 2.5: Independent dual-sensor scanning for real-time image-based autofocusing [56]

2.2.3 Split-beam Array Method

In the aforementioned dual-sensor independent scanning approach, multiple images are captured to compute the focus position as the sample moves across different focal planes. Virag *et al.* introduced a beam-splitting array method, which simultaneously captures images of different focal planes on a single image sensor [82]. Figure 2.6 represents the imaging principle of the system, where the quasi-focal optical components consist of the primary imaging camera and an auxiliary quasi-focal camera. The beam-splitter array serves to separate the light beams and reflect them onto different areas of the quasi-focal sensor, allowing the system to simultaneously capture images on multiple focal planes. By selecting a 45° semi-reflective surface within the beam-splitting array method, one can ensure that all the beams reflected off the surface possess approximately equal intensity. With the images captured by the quasi-focal sensor, an optimal quasi-focal position can be inferred using specific quasi-focal metrics and fitting models.

2.2.4 Tilted Sensor Approach

The tilted sensor method employs a tilted quasi-focal sensor to image the oblique cross-section of a sample. The optimal focal position can be inferred by real-time pinpointing of the peak value on the contrast curve. Philips and Leica have further refined and developed this original concept, and the tilted sensor technique has now become one of the autofocusing technologies widely adopted in commercial WSI systems [95, 29, 81].

Figure 2.7 illustrates the principle and operational process of the tilted sensor concept. In Figure 2.7(a), the quasi-focal sensor is tilted at an angle relative to the

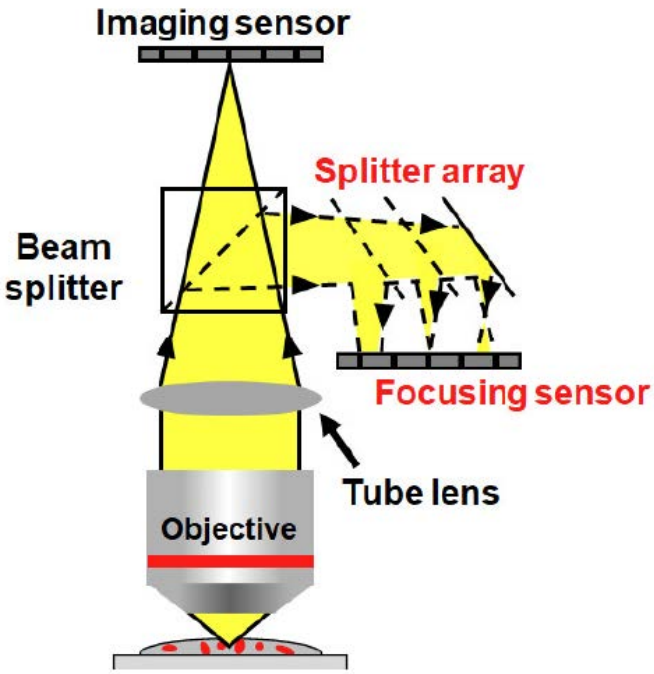

Figure 2.6: Beam splitter array for real-time image-based autofocusing [82]

in-focus plane. This quasi-focal sensor can be a 2D area sensor or a 1D linear sensor. The overlapping position between the quasi-focal sensor and the in-focus plane is referred to as the co-focus point in Figure 2.7(b). The focusing range is determined by the z-range; the greater the tilt angle, the longer the focusing range. During the scanning process, both sensors capture images of the sample. For each pixel of the captured data, a contrast value can be determined based on surrounding pixel values. Then, by dividing the contrast values of the quasi-focal sensor by those of the imaging sensor, a contrast curve is obtained, as depicted in Figure 2.7(c). The peak of the contrast curve identifies the pixel with the highest contrast value, i.e., the pixel at the optimal focal position. The co-focus point can also be plotted on this contrast curve. In Figure 2.7(c), the pixel distance between the co-focus point and the peak

of the contrast curve represents a physical distance along the z-axis. This distance indicates the offset between the current position of the objective lens and its optimal focal position - representing how much the objective lens needs to move axially to achieve optimal focus. When the imaging sensor is centered on the objective's field of view, the quasi-focal sensor can be offset from the center of the optical field of view. The quasi-focal sensor detects image data before the imaging sensor detects the same area. Similarly, volume cameras comprised of multiple linear CCDs coupled with optical fibers can be arranged at tilted angles for autofocusing [65]. Bravo and colleagues reported on using nine sensors coupled with fibers to capture images on different focal planes, facilitating real-time image-based autofocusing [9].

2.2.5 Phase Detection Method

Phase-detection autofocusing has been extensively adopted in the majority of digital single-lens reflex cameras (DSLRs). This technique typically works by splitting the incoming light into a pair of images. The distance between these two images is then measured, allowing for an inference of the focal offset. Here, the term phase pertains to the translational offset between the two images (or phase shift in the Fourier domain). Inspired by the phase detection concept in photography, an autofocusing attachment kit has been developed to facilitate full scanning imaging using a standard microscope [23]. As illustrated in Figure 2.8(a), two aperture-modulated cameras are attached to the eyepiece for phase detection autofocusing. By adjusting the positions of the two apertures, the viewpoints can be effectively altered via both eyepiece outlets. When the specimen is placed at the focal point, the images captured by both cameras will be identical. If the specimen is positioned off the focal point,

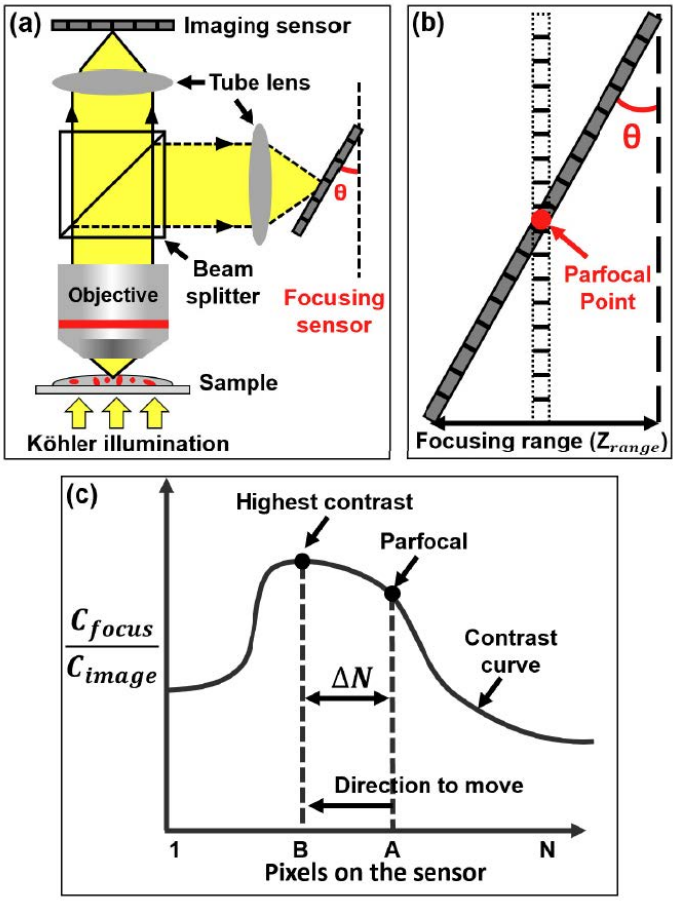

Figure 2.7: Beam splitter array for real-time image-based autofocusing [95, 29, 81]

it will project at two distinct angles, leading to a translational offset between the captured images. This offset is directly proportional to the defocusing distance of the specimen. Hence, by identifying the translational offset of the two captured images using phase correlation, the specimen's optimal focal position can be retrieved without necessitating a z-axis scan.

Figure 2.8(b) showcases another autofocusing scheme grounded in the phase detection concept [41]. A dual-aperture mask is positioned on the pupil plane to modulate the specimen's light. Unlike the method employing two aperture-modulated cameras,

here, only a single focus sensor is utilized to capture the image modulated by the dual-aperture mask. In this scenario, the image acquired from the focus sensor encompasses two replicas of the specimen, with the translational offset between them being directly proportional to the defocus distance. Figure 2.8(b) presents the raw image captured by the focus sensor, where duplicates of the specimen are discernible. The distance between these two duplicates can be recovered through the auto-correlation analysis depicted in Figure 2.8(b). Figure 2.8(c) displays a similar phase detection scheme proposed by Silvestri *et al.* [77]. Analogous to the dual-aperture modulation method, only a single camera is employed for focusing. A wedge plate is inserted into the pupil plane, directing half of the light beam at a slightly inclined angle. Consequently, the image acquired from the focus sensor contains two replicas of the specimen separated by a definite distance. The defocus distance can be inferred from the translational offset between the two replicas. For the configurations shown in Figure 2.8(a) and (b), aperture masks are used to limit the light on the pupil plane, offering them a relatively longer autofocusing range. In contrast, the system in Figure 2.8(c) has a shorter autofocusing range. The use of the dual-aperture mask doesn't impede its application in fluorescence microscopy. A beam splitter can be employed to guide the intense excitation light through the dual-aperture mask, enabling camera detection of the specimen's weak fluorescence emission.

2.2.6 Dual-LED Illumination Technique

The dual-LED illumination method has been proven to achieve single-frame autofocusing even when the sample is in continuous motion [43, 44, 42, 31, 22]. Figure 2.9(a) demonstrates one such configuration where two near-infrared LEDs are positioned at

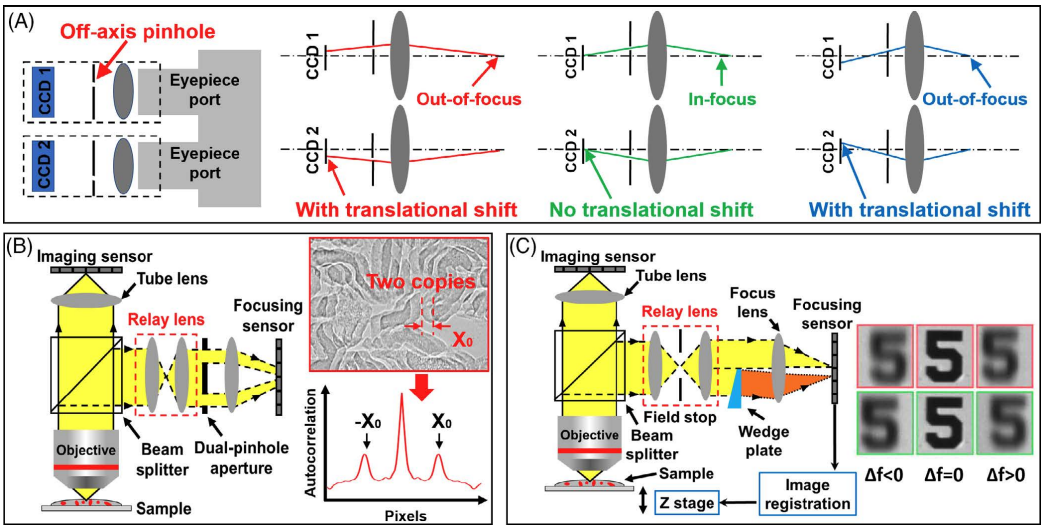

Figure 2.8: Phase detection for real-time image-based autofocusing [23]

the post-focal point of a condenser lens for sample illumination. These two LEDs illuminate the sample from two different angles of incidence and can be regarded as spatially coherent light sources. Using a hot mirror, the near-infrared light can be directed to the quasi-focal sensor shown in Figure 2.9(a). Consequently, the images captured by this quasi-focal sensor will contain replicas of the sample images that are spaced a certain distance apart. Specifically, the quasi-focal sensor is positioned at a predefined offset distance relative to the imaging sensor. When the sample is at the quasi-focal position, the image captured by this sensor will still contain two replica images of the sample outline. Similar to the dual-pinhole template method, the interval between the two image replicas can be determined through auto-correlation analysis, thus recovering the defocus distance. The preset offset in Figure 2.9(a) is configured to enhance the accuracy of the auto-correlation analysis and generate defocus contrast for transparent samples. If the direction of sample movement is perpendicular to the translation direction, autofocusing can be achieved even with

continuous sample movement. This dual-LED approach has also been demonstrated to measure the focus plane using only the primary camera.

Figure 2.9(b) illustrates the dual-LED method using color multiplexed illumination. In this setup, a color LED array is employed for sample illumination. For regular bright-field image acquisition, all LEDs are turned on, as shown on the left side of Figure 2.9(b1). Between two bright-field acquisitions, red and green LEDs are activated for multicolor illumination. If the sample is placed out of focus, the red and green image replicas will be separated by a certain distance, as depicted in Figure 2.9(b1). Subsequently, the translation between the red and green image channels can be identified by maximizing image mutual information or cross-correlation. The resulting translation is utilized for dynamic focus correction during scanning. Figure 2.9(b2) displays the WSI method for dual-LED autofocusing based on color multiplexing.

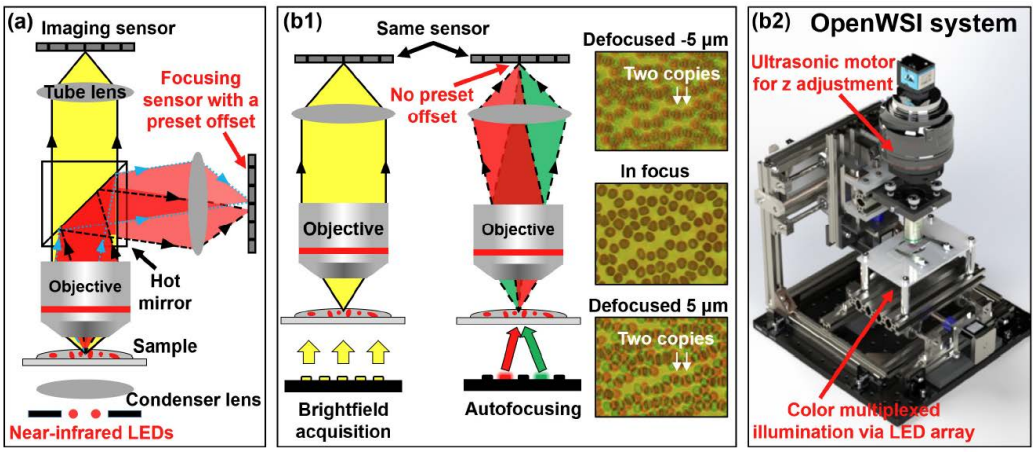

Figure 2.9: Dual-LED illumination for single-frame autofocusing [43, 44, 42, 31, 22]

2.3 Deep Learning-based Autofocusing Techniques

With the rapid advancements in the fields of artificial intelligence and computer vision, research directions that integrate cutting-edge AI algorithms with microscopic imaging techniques have garnered significant attention from researchers [16, 32, 11, 13, 12, 6, 93, 45, 39, 89, 14, 63]. Currently, there are two main methods for implementing autofocusing in microscopic imaging using deep learning: focus estimation and focal plane recovery. In the focus estimation method, neural networks are typically used to learn the mapping relationship between defocus images to the defocus distance. Once the defocus distance is obtained, mechanical movement compensates for the defocus distance to capture the true in-focus image. In the focal plane recovery method, a neural network is constructed to learn the inverse imaging process. The input of defocus images is processed through the network, directly outputting the recovered in-focus image. The focus estimation method is suitable for scenarios where authentic captured images are required, while the focal plane recovery method is applicable to high-speed scanning, offline processing, and other scenarios, achieving virtual in-focus imaging.

2.3.1 Focus Prediction Methods

Jiang *et al.* were the first to utilize artificial intelligence techniques to quickly achieve autofocusing on a single frame image [30]. Through multi-domain learning (spatial, frequency, and multi-domain), this method can capture and learn focus-related information from various domains (different imaging conditions, optical setups, or tissue types) as shown in Figure 2.10. This cross-domain learning approach enhances the

system's robustness and adaptability, improving imaging speed and quality even under complex imaging conditions. The in-focus distance is estimated through a neural network, and then the mechanical platform is adjusted to compensate for this distance, achieving efficient autofocusing. However, this method did not fully take into account the inherent limitations of imaging.

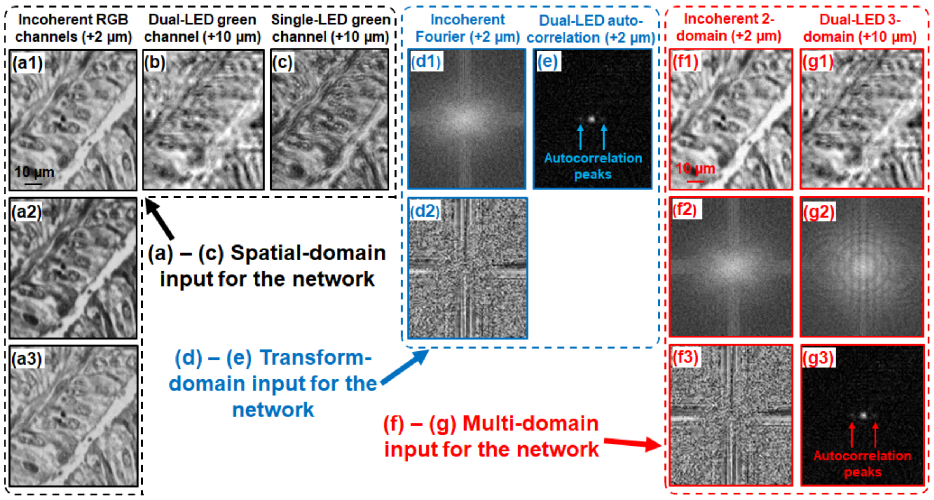

Figure 2.10: WSI autofocusing method based on deep learning (Focus Prediction) [30]

Tathagato *et al.* utilized lightweight network designs like MobileNet_v2 to create an autofocusing network [15]. They used the difference between two defocus images taken at fixed intervals as the network input, and the output was the estimated in-focus distance, as shown in Figure 2.11. By utilizing the defocus difference design, image details related to the distance were retained, avoiding the influence of sample diversity on prediction results. However, this method requires two images to estimate the in-focus distance, whereas Jiang's method [30] only needs a single image. Scanning additional images for exposure results in extra scanning time costs during the focus image construction process.

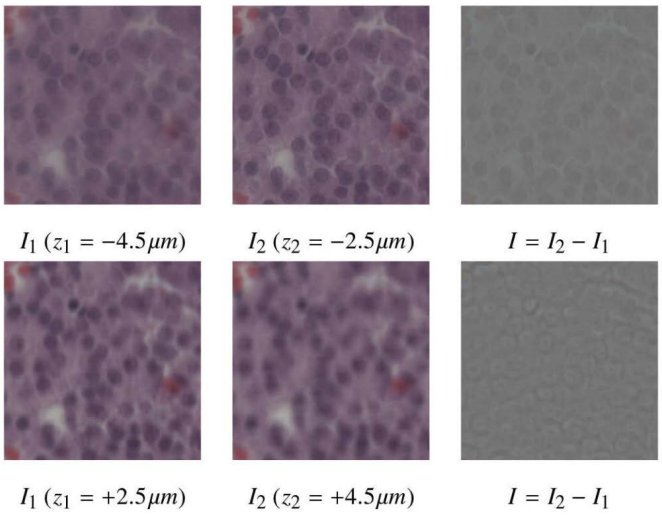

Figure 2.11: WSI autofocusing method based on deep learning by two images (Focus Prediction) [15]

PINKARD *et al.* proposed a method for single-shot autofocusing microscopy imaging under coherent light using deep learning techniques [64], as depicted in Figure 2.12. Traditional autofocusing techniques typically rely on continuous adjustments and verifications to locate the optimal focal plane. This process can be both time-consuming and potentially imprecise, especially when dealing with complex sample structures or a wide range of focus. The method aims to address a core challenge in microscopy imaging: how to quickly and accurately focus on the optimal plane of the sample without multiple scans or adjustments. This approach utilizes convolutional neural networks in microscopy imaging by training a model to identify and predict the best focus position. During the training phase, the in-focus image Ground Truth (GT) is determined by identifying the maximum spectral energy from the non-coherent z-stack. The input images, taken using coherent illumination of defocus images, undergo a frequency domain transformation before being fed into the network. The network's output is an estimated focal length. The paper provides a

detailed description of the entire microscopy imaging system, including how to integrate the deep learning model, hardware configurations, and associated software tools. The authors conducted a series of experiments to validate the method's effectiveness, comparing it to traditional techniques. These experiments covered various types of biological samples, demonstrating the method's versatility across different scenarios. While this method offers rapid network processing speeds, it requires optical modulation of the WSI imaging system (using multiple illumination methods) and performs suboptimally in terms of sample diversity.

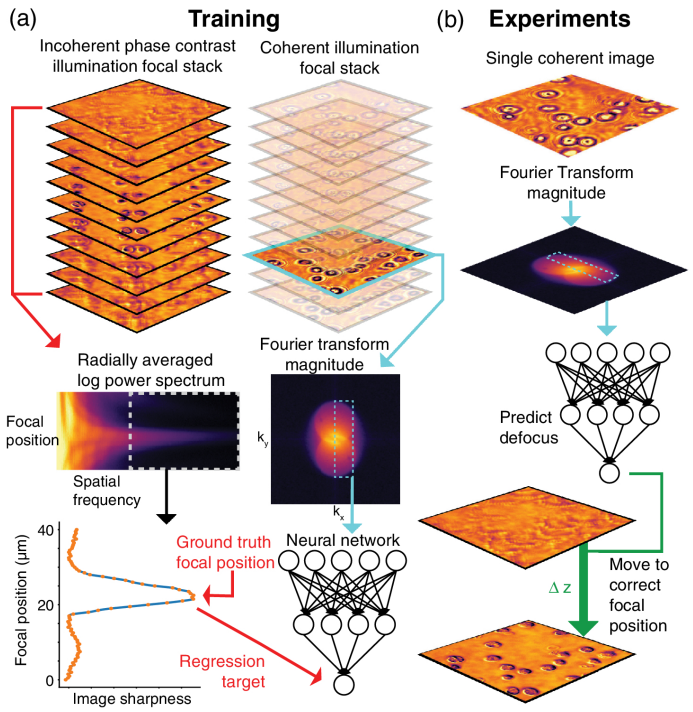

Figure 2.12: Single-shot autofocusing microscopy by deep learning (Focus Prediction) [64]

Lee *et al.* introduced a dual-network method for automatic focusing in Scanning

Electron Microscope (SEM) imaging based on deep learning [37]. This method consists of an Auto-Focus Evaluation Network (AENet) and an Auto-Focus Control Network (ACNet), as illustrated in Figure 2.13. AENet evaluates image quality based on the current image and another image with the same dimensions representing normalized magnification values, given a specific working distance. The effective utilization of ACNet requires the integration of AENet scores, SEM parameters (like working distance and magnification), and traditional image quality indicators (such as image variance and entropy). Subsequently, the adjusted working distance value is relayed back to the SEM. AENet is designed to assess the quality of a given image with a score range from 0 to 9. ACNet can precisely control the SEM focus online, based on AENet's output, for any lateral sample position and magnification. While this dual-network approach demonstrated promising autofocusing performance on three training samples, the workflow of the dual networks operates in a feedback manner, making joint optimization relatively intricate.

Tang *et al.* proposed a strategy based on the Deep Image Prior (DIP), embedding neural networks within physical models [80]. This is designed to approximate processes that are challenging to model or parameters difficult to measure, aiming to find the optimal solution in single-variable optimization problems, as shown in Figure 2.14. The extensive training, large sets of manually labeled data, and limited generalization have constrained the application of deep neural networks under supervised learning. Methods rooted in neural networks necessitate substantial data to fit physical models or deduce inverse relations. This approach is cumbersome and time-consuming since some phenomena can be precisely simulated and analyzed in optics. Moreover, there's

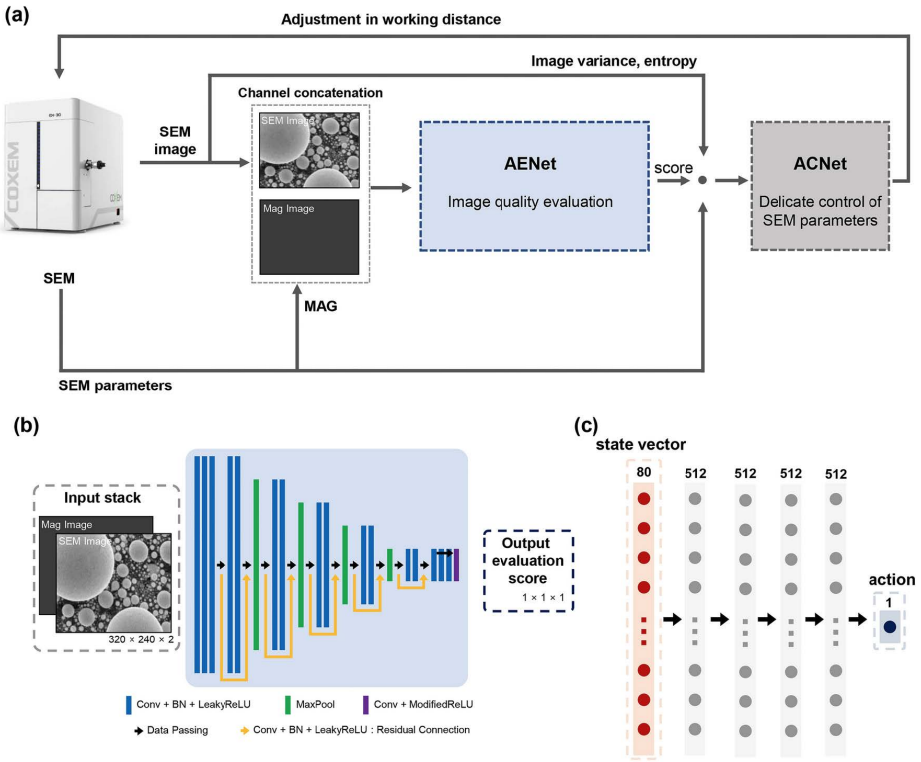

Figure 2.13: Architectures of the autofocusing SEM based on a dual deep learning network (Focus Prediction) [37]

a risk that neural network-based methods may not align with physical realities, leading to inevitable readjustments. Tang introduced an Untrained Physical Network (UPN) that predicts diffraction distances solely from a known phase object's diffraction pattern. Experimental results demonstrated that UPN could consistently and accurately predict distances associated with different targets, diffraction distances, and phase ranges while requiring only a brief training period. Furthermore, once trained, the UPN can generalize to other targets as long as the actual diffraction process remains unchanged. Compared to autofocusing metrics of holographic reconstruction and traversal methods, UPN boasts advantages in both speed and precision. It also exhibits commendable noise resistance, which is meaningful for autofocusing

in holographic reconstruction and imaging. However, this method still relies on the availability of accurate physical models and may have certain limitations in situations without analytical expressions.

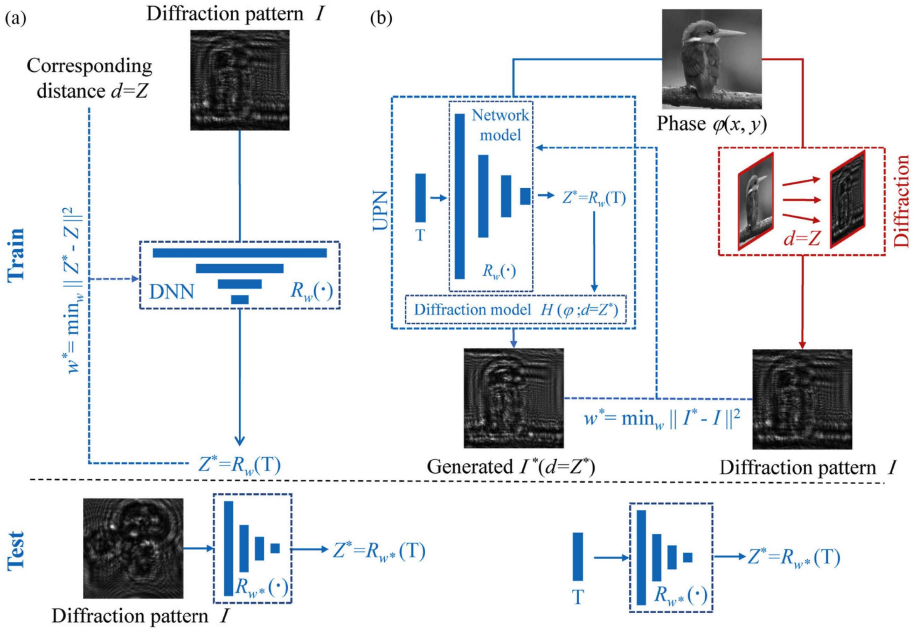

Figure 2.14: Schematic illustration of the UPN network in prediction of diffraction distance (Focus Prediction) [80]

Montoya *et al.* introduced a regression model based on convolutional neural networks (CNN), named FocusNET, designed to predict the accurate reconstruction distance of original holograms in Digital Lens-free Holographic Microscopy (DLHM) [57], as depicted in Figure 2.15. In Digital Holographic Microscopy (DHM), a significant challenge lies in determining the precise location of a sample within the inspection volume without any supplementary procedures. For weakly scattering specimens containing axially disconnected samples, digital holograms provide plane-by-plane information about the entire volume. However, there isn't a direct method to ascertain the reconstructed focal plane. Montoya presented a physico-mathematical formula and

extended its application to DLHM setups that differ from the optical and geometric conditions used during the recording of the training dataset. By applying this method to holograms of various samples recorded using different DLHM configurations, tests validated its distinctive feature. Moreover, the study also furnished a comparison of FocusNET with conventional autofocusing techniques in terms of processing time and accuracy. Compared to methods that utilize a series of reconstructions to locate the optimal focal plane, FocusNET's performance is accelerated by a factor of 600, primarily because it eliminates the need for hologram reconstruction.

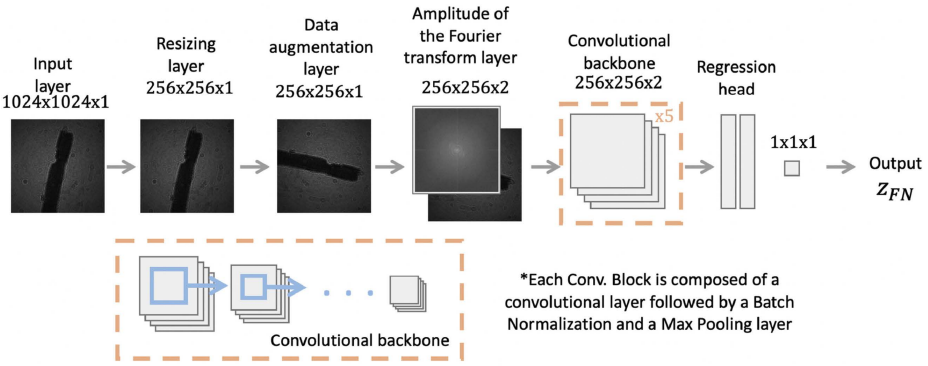

Figure 2.15: DLHM FocusNET architecture (Focus Prediction) [57]

2.3.2 In-focus Restoration Methods

Wu *et al.* harnessed deep learning techniques to achieve three-dimensional virtual refocusing of fluorescence microscopy imaging [87], offering an efficient means to refocus captured fluorescence microscopic images, resulting in clearer and more accurate three-dimensional structures, as shown in Figure 2.16. Fluorescence microscopy is commonly employed to observe cells and intracellular molecular structures. However, acquiring sharp three-dimensional images often necessitates multiple scans and post-processing, which can be time-consuming and resource-intensive. By leveraging

neural networks, they demonstrated how a clear three-dimensional structure can be inferred from a single fluorescence microscopic image. This technique facilitates the virtual refocusing of an individual image, eliminating the need for multiple scans. To train and validate their approach, multiple fluorescence microscopy datasets were utilized, undergoing preprocessing and augmentation to fit the deep learning models. Compared to traditional three-dimensional refocusing techniques, this deep learning approach yielded faster and higher-quality outcomes. This method holds vast potential for biomedical research, offering researchers an efficient tool to explore cellular and intracellular structures. It represents an innovative deep learning approach in the realm of fluorescence microscopy imaging, elevating the capability of three-dimensional virtual refocusing to new heights and marking a significant advancement in this domain. However, compared to structurally simple fluorescence images, pathological microscopy images possess more intricate biological structure features and require higher imaging quality and efficiency.

Gan *et al.* introduced a rapid and accurate deep learning-based autofocus method, addressing the challenge of focus instability in Light Sheet Fluorescence Microscopy (LSFM) [18], as depicted in Figure 2.17. LSFM, recognized as a promising tool in biological research due to its capability to continuously observe live cell dynamics for hours and days, places stringent demands on the light sheet and the detection focal plane to achieve optimal image quality. Spatial light modulators can generate light sheets, modulating the excitation beam into multi-depth lattice patterns for multiplexed structured illumination. Defocusing information is encoded into combinations of distinct stripe patterns of different depths. Concurrently, neural networks can be

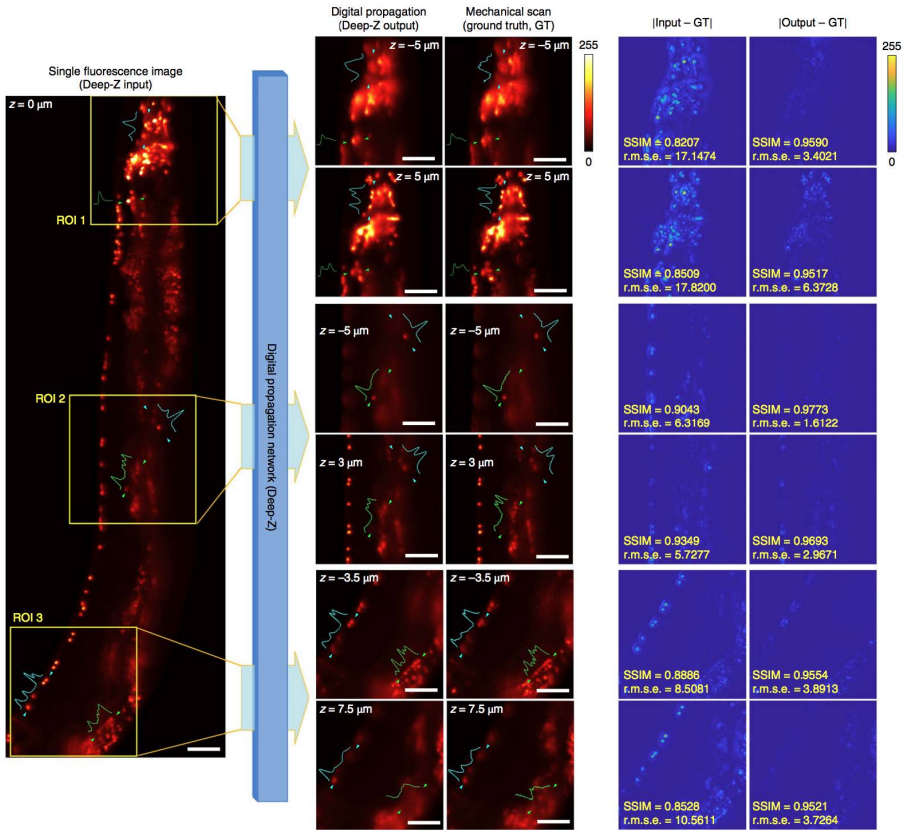

Figure 2.16: ROIs are refocused using Deep-Z to different planes within the sample volume (In-focus Restoration) [87]

employed for high-precision decoding or predicting the defocus amount. The network architecture adopted by Gan is memory-efficient, demands a minimal training dataset, and is easily adaptable to various experimental conditions. The method is compatible with any light sheet imaging apparatus equipped with a spatial light modulator for light sheet generation. The proposed neural network architecture boasts commendable generalizability benefits for untrained sample types. However, the approach requires light modulators and other light sheet generation devices, making it relatively costly.

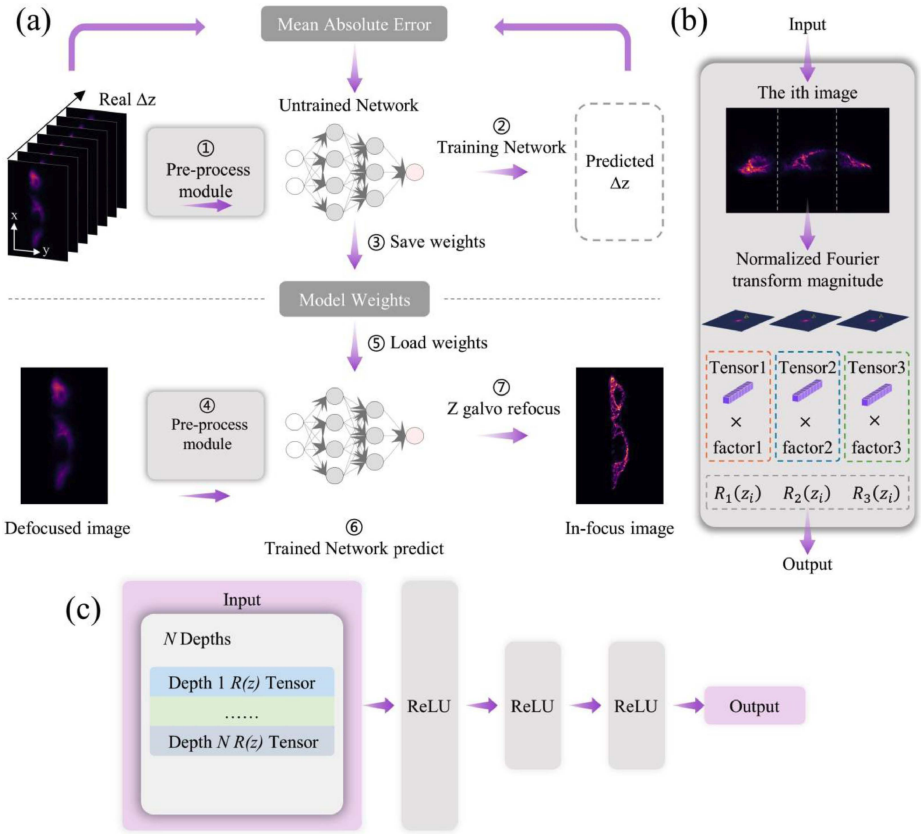

Figure 2.17: Architectures of the autofocusing LSFM based on a multiplexed structured illumination network (In-focus Restoration) [18]

Huang *et al.* proposed a phase-recovery method based on a Convolutional Recurrent Neural Network (RNN) [28]. This technique rapidly reconstructs phase and amplitude information on samples using multiple holograms captured at varying distances from the sample to the sensor. It also accomplishes autofocus within the same network, as depicted in Figure 2.18. Digital holography is among the widely used label-free imaging techniques in biomedical imaging, and recovering lost phase information from the hologram is a crucial step in holographic image reconstruction. Huang introduced a deep learning-based holographic image reconstruction and phase retrieval algorithm, trained using a Generative Adversarial Network (GAN). This

imaging framework uses multiple input holograms, which are back-propagated with zero-phase to a common axial plane, achieving autofocus and phase retrieval at its output simultaneously. By employing dilated convolution kernels, there's no need for any spatial back-propagation steps. The captured original holograms of the object are directly fed into the trained RNN, with the focused image reconstruction done at its output. The efficacy of this deep learning-based holographic imaging method was validated by imaging microscopic features of human tissue samples and Gram-stained smears. Compared to existing methods, the proposed approach enhances the quality of the reconstructed images while also improving the depth of field and inference speed. However, holographic imaging lacks an objective lens, resulting in holograms distinctly different from pathology slide images.

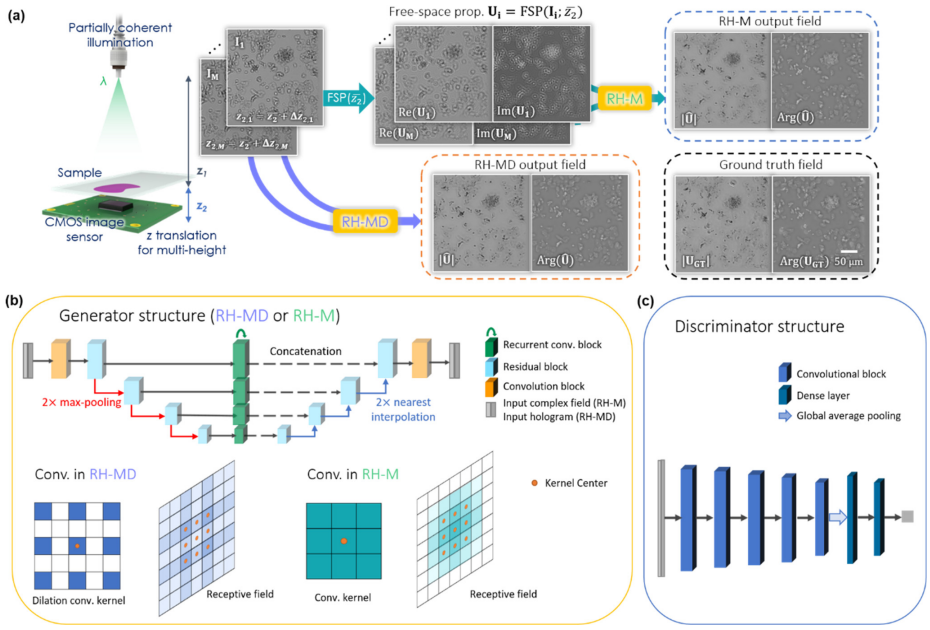

Figure 2.18: Recurrent holographic imaging framework (In-focus Restoration) [28]

Xu *et al.* introduced a deep learning-based image processing technique to obtain autofocusing images from Surface Plasmon Resonance Microscopy (SPRM) without

adding complexity to the optical system [88], as illustrated in Figure 2.19. SPRM inevitably suffers from non-uniformities and shifts in focus, particularly during prolonged recordings, resulting in image distortion and inaccurate quantification. Traditional focus correction methods necessitate additional optical components to detect and adjust focus conditions. While digital holographic image processing algorithms can, in principle, reconstruct images on any focal plane, they grapple with challenges like twin-image interference, missing initial phase, and unknown object positions. Xu trained a network model using thousands of SPRM images of nanoparticles acquired at different focal lengths for correcting focus drifts in SPRM. The trained model is capable of generating in-focus SPRM images directly from a single defocused image, without knowledge of the focal condition during recording. A GAN model was constructed and trained using thousands of SPRM images acquired at various focal planes. This trained model automatically corrects the focus of the input SPRM image and provides a refocused image at the output. The methodology was experimentally studied by monitoring nanoparticles in both static and dynamic settings and quantitatively compared to assess its efficacy. Experiments demonstrated the method's effectiveness in both static and time-lapse monitoring. Hence, the proposed autofocusing technique offers an effective approach for enhancing the consistency of SPRM research and long-term monitoring. However, GAN networks are typically challenging to train and inevitably produce artifacts and noise.

2.4 Scanning Strategy

Whole slide digital pathology imaging is a medical imaging technique that has garnered significant attention in recent years. It enables pathologists to view, analyze,

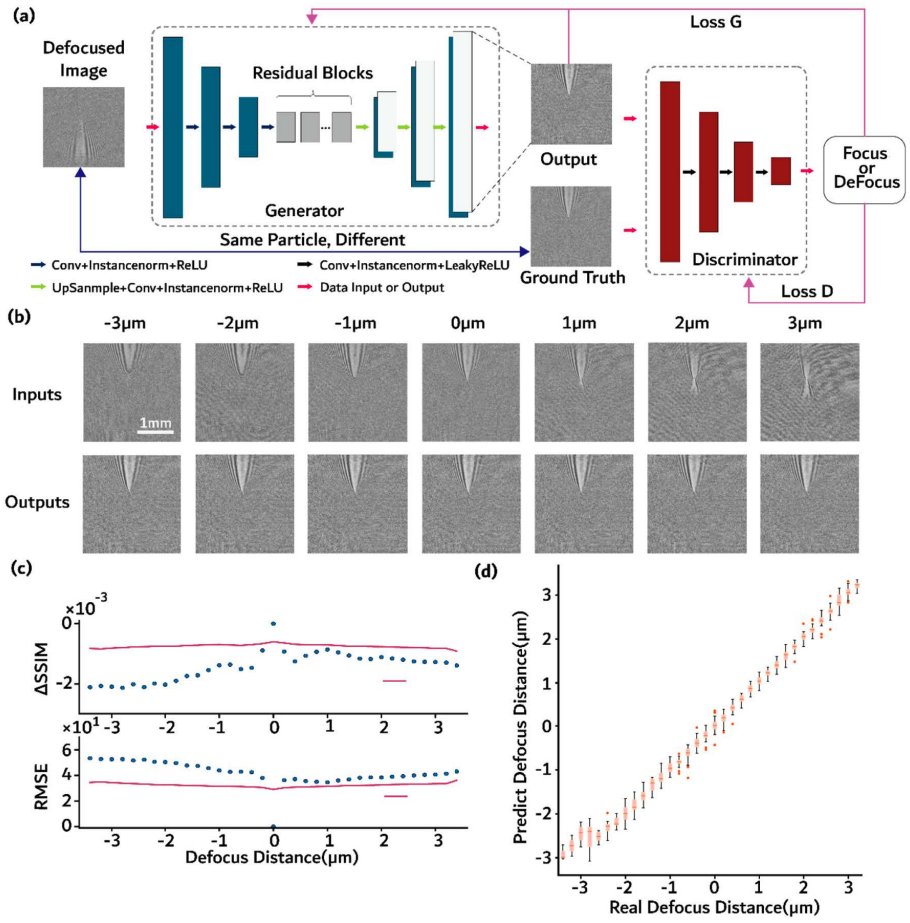

Figure 2.19: GAN models and particle-by-particle correction (In-focus Restoration) [88]

and interpret tissue slice images on computer screens. To acquire high-resolution digital images, scanning methods are typically employed. We introduce two prevalent scanning techniques: tile scanning and line scanning methods. The scanning strategy is based on the focus map surveying method, illustrated in Figure 2-21.

2.4.1 Tile Scanning and Line Scanning Methods

Tile scanning, often referred to as regional or block scanning, stands as a foundational method adopted by the vast majority of pathology slide scanners. It involves decomposing the slide into several small tiles or regions, scanning each tile along the x-y axes, and focusing using axial movement in the z-direction. Upon completion of the scan, all the tiles are reassembled to form a comprehensive digital pathology slide image. This approach is particularly suitable for larger pathology slides or samples requiring scanning at high resolutions. Given its focus on processing smaller tiles, it effectively allocates computational resources. A potential drawback of this method arises at tile boundaries where suboptimal stitching may occur. Thus, high-quality image-stitching algorithms are essential to ensure image continuity.

Linear scanning, on the other hand, involves scanning the slide in continuous lines or paths, mirroring the operations of conventional scanners or printers. The scanning head moves linearly along a predetermined path, capturing images along this trajectory, for example, solely in the x-direction. This method yields continuous, seamless images, eliminating concerns about stitching between tiles. However, in the linear scanning approach, ensuring high-quality image capture necessitates that the CCD sensor's signal and the movement of the scanned slide sample be strictly synchronized. Such a mechanism implies that real-time previews of specific slide image details are challenging to achieve during the scanning process. Additionally, the precision demanded by linear scanning for mechanical and control systems results in a relatively higher implementation cost.

2.4.2 Focus Map Surveying Method

High-resolution, in-focus images of entire slide specimens can be achieved by repeatedly applying the z-stack auto-focusing process to each tile. However, the auto-focusing process may entail a significant amount of time capturing z-stacks at multiple locations. Assuming images are captured at a rate of 20 frames per second, scanning five distinct focal points would require 0.25 seconds for each tile. As a result, an image comprising 500 tiles might take up to 150 seconds to capture, excluding the time for deceleration, acceleration, and positioning to move the slide to different lateral and longitudinal positions. Applying conventional image-based focal measurement methods for auto-focusing on each tile is not the most efficient solution. To reduce time costs, many WSI systems either create a focus map before scanning or conduct a focal scan for every certain number of tiles or lines. The number and positioning of focal points are usually determined by the user.

Figure 2.20(a) illustrates the process of generating a focus map by the mapping method [7]. Initially, the system selects focal points based on the sample's characteristics, distributing them evenly across the entire slide. Each focal point employs triangulation to produce a focus map of the tissue surface, subsequently filling the vacant areas. Triangulation stands as a typical method for focus map generation, as shown in Figure 2.20(b). Linear scanning methods generally offer superior auto-focusing performance compared to traditional 2D tiles since linear sensors can adjust focus at shorter intervals. Another approach to creating a focus map involves auto-focusing every nth tile, referred to as skip-tile in Figure 2.20(b). Here, it's assumed that the focus is shared between tiles. However, compared to the focus map method, its in-focus performance is subpar and might include more out-of-focus areas. Yet,

the skip-tile method doesn't necessitate returning to specific axial positions with sub-micron precision. Its demands for motion repeatability are less stringent compared to the focus map method. Nonetheless, for both methods, adding more focal points can enhance the overall focusing performance's accuracy but comes at the cost of increased auto-focusing time.

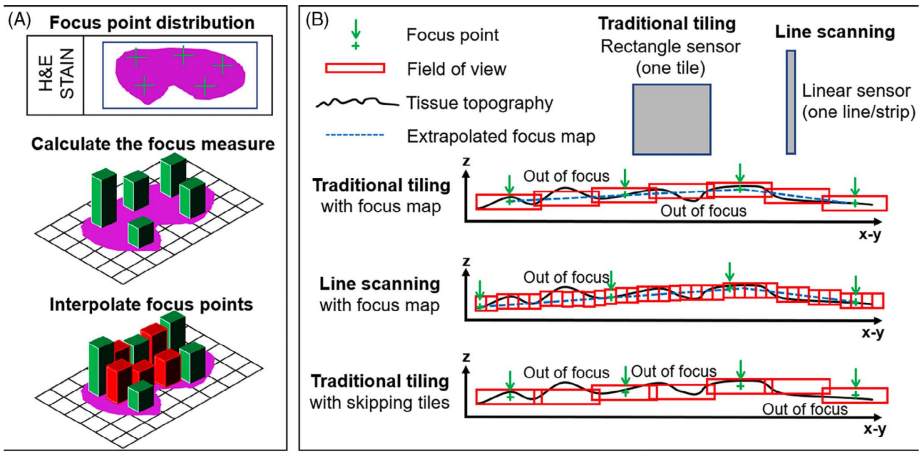

Figure 2.20: Focus map generation and scanning methods [7]

2.5 Conclusions

This chapter provides an overview of the current state of autofocusing research. It encompasses several representative methods for microscopic imaging autofocusing, including reflective hardware-based autofocusing, real-time image-based autofocusing, and deep learning-based autofocusing. From the review, it is evident that deep learning-based autofocusing methods have been extensively applied in various domains in recent years. They offer numerous advantages, such as high accuracy, rapid processing speed, robust generalization capabilities, reduced dependency on hardware, and minimization of human-related variables. Deep learning-based autofocusing

techniques, like focus prediction and focal plane recovery methods, have progressively become the predominant strategies for microscopic imaging autofocusing.

Chapter 3

Aberration-aware Focal Distance Prediction

3.1 Introduction

WSI is an essential technology for digital pathology, the performance of which is primarily affected by the autofocusing process. Conventional autofocusing methods either are time-consuming or require additional hardware and thus are not compatible with the current WSI systems, as mentioned in Chapter 1.1. Compared to mechanically adjusting the focal distance on a tile-by-tile basis, using advanced machine learning algorithms to predict focus position of pathological images is an efficient approach. In the current deep learning-based focus-prediction autofocusing methods [30, 15, 66], all images are treated by the same neural network to derive the defocus distance. Howerver, as a practical optical system, the effect of optical aberrations inevitably exists in WSI. The images with positive / negative defocus are not symmetric with respect to the focal plane, resulting in images with different levels of defocus

artifacts, as illustrated in Figure 3.1. Therefore, ignoring the undesirable effect of optical aberrations, the deep model would be at the risk of overfitting.

Inspired by this physics-based observation, in this paper, we consider two interacted issues jointly for autofocusing: 1) how to reduce the effect of optical aberrations effectively; 2) how to determine the defocus distance accurately. For the first issue, we propose a defocusing classification network, which can determine images with either positive or negative defocus offset. By classification, samples within the same category share similar appearance characteristics, which remedies the undesirable effect of optical aberrations. For the second issue, we propose a two-branch refocusing network, which includes two CNN models for estimating defocus distance, one for images with positive defocus offset, and the other for images with negative defocus offset. Experimental results demonstrate that our method achieves superior autofocusing performance compared with the state-of-the-art (SOTA).

3.2 Preliminaries and Motivations

In this section, we introduce related preliminaries, including DoF and defocus definition, the effect of optical aberrations and defocus images in WSI, which serve as the motivations of modules of our proposed method.

3.2.1 Depth of Field and Defocus Definition

In microscopy, the DoF is determined by the distance between the focal plane and the farthest plane where the captured image is still clear. Mathematically, DoF is

determined by the numerical aperture (NA) of the objective lens

$$DoF = \frac{\lambda \cdot m}{NA^2} + \frac{m}{M \cdot NA} e, \tag{3.2.1}$$

where λ is the wavelength of illumination, m is the refractive index, e is the pixel size of detector and the lateral magnification of microscope objective is M. In WSI, NA of the objective lens is high. Thus DoF is usually small (lower than $1\mu m$).

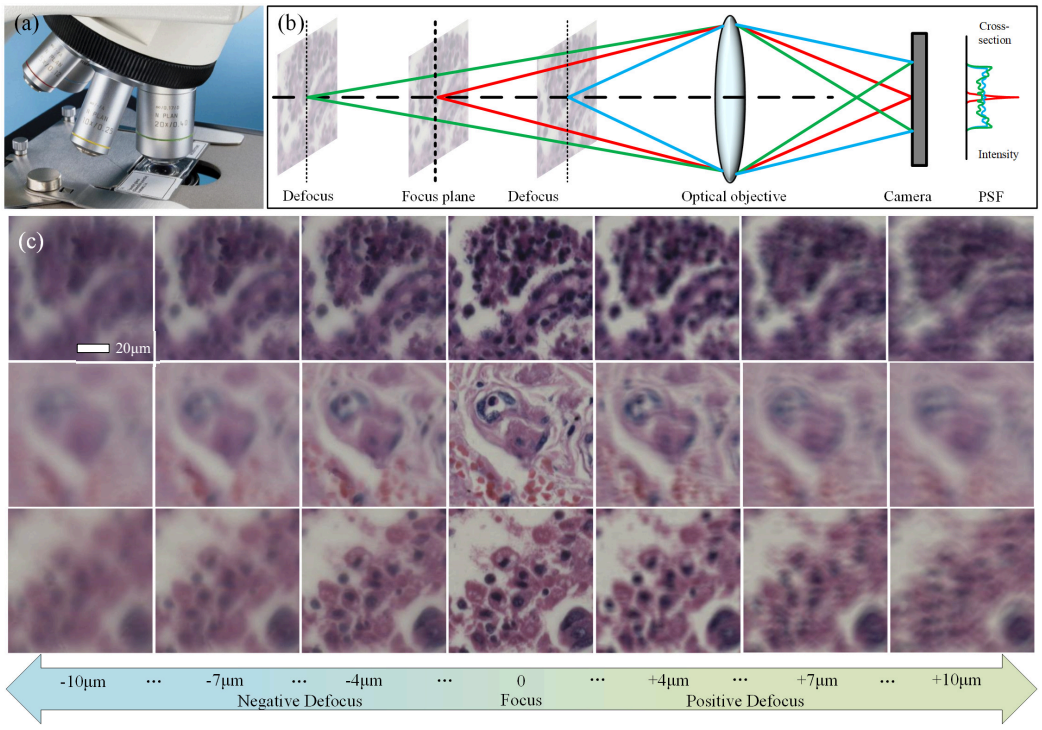

Figure 3.1: (a) The microscope system of WSI. (b) Defocus and focusing model. The PSF is the cross section of image intensity. (c) Illustration of the asymmetric effect of optical aberrations on three samples. The defocus distances are -10μm, -7μm, -4μm, 0μm, 4μm, 7μm and 10μm, respectively.

The imaging model in WSI is illustrated in Fig. 3.1. Fig. 3.1-(a) illustrates the microscope system of WSI, where the objective lens is placed above the sample and

is moved along the axial direction to adjust the focus of the microscope. The optical model is shown in Fig. 3.1-(b), which indicates that the effect of defocus would result in a blurred image on the camera. As an example, we exhibit images of three samples at different defocus distances along the optical axis, as shown in Fig. 3.1-(c). It can be seen that a clear sample image is captured in the focal plane, while images become blurring when their locations deviate from the focal plane.

It is worth noting that, in practice, to capture clear sample images, it is not necessary to move samples to precisely locate in the focal plane; clear images can be obtained as long as the focusing errors are within the range of DoF of the objective lens. Thus, we define images captured out of DoF as the defocus ones.

3.2.2 Optical Aberrations

An ideal imaging model is shown in Fig. 3.2-(a), where a very thin lens is used. However, the lenses used in practice are all with the thickness of a certain degree. Thus the captured images would suffer from the effect of optical aberrations, as shown in Fig. 3.2-(b). By comparing Fig. 3.2-(a) and (b), it can be seen that, if both are used to capture a very thin sample, in ideal imaging model, images on opposite sides of the focus are symmetric with respect to the focal plane; while in real imaging model, images with positive / negative defocus are asymmetric with respect to the focal plane. This physical observation is the primary motivation of this work. Specifically, the optical aberrations of WSI are mainly due to spherical aberration. In the following, we give a detailed interpretation about this effect.

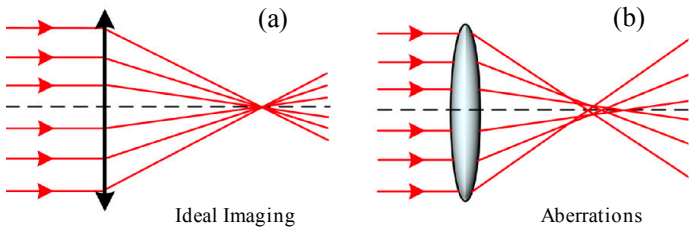

Figure 3.2: The model of single lens imaging. (a) Ideal focusing model. (b) Practical model with optical aberrations.

Spherical Aberration In Microscopy

To achieve high level of optical performance, it is necessary to design and manufacture the single lens carefully, which is usually the spherical surface since it is easier to fabricate than non-spherical curved surfaces. Moreover, considering the limitation of a single spherical lens in focus ability, multiple lens elements (*e.g.* the objective lens of microscopes) are assembled for image shooting, which must be precisely located along the optical axis in order to balance the optical aberrations. However, this balance can be upended due to the refractive index mismatch caused by transmission media, cover glass, or the specimen itself. Light rays that approach the focus at a larger angle experience greater refraction at an interface. It leads to spherical aberration, *i.e.*, the focus position differs in depth between the central and peripheral light rays, as illustrated in Fig. 3.3-(a).

In WSI, air and tissue sample are involved in refraction, whose refractive indexes are $n_1 = 1$ and $n_2 = 1.35 \sim 1.55$, respectively. In the focus scenario, as shown in Fig. 3.3-(b), light rays from the objective lens are concentrated on the air-tissue interface. When the objective lens is brought closer or farther to the tissue, as illustrated in Fig. 3.3-(c) and (d), the asymmetric effect of spherical aberration generates due to the refractive index $n_1 \neq n_2$, resulting in different defocus artifacts. They are called

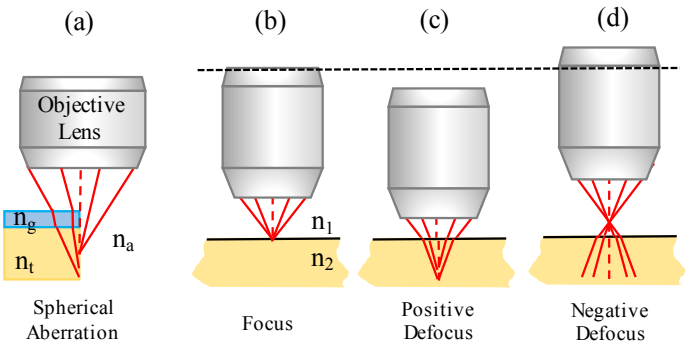

Figure 3.3: The spherical aberration phenomenon in microscopy. (a) Right: ideal focusing in the air; Left: The fact with spherical aberration caused by the refractive index mismatch. n_a, n_b, n_c stand for the refractive index of air, cover glass and cell tissue. (b) Simplified focusing model only with two kinds of transmission medium, air n_1 and cell tissue n_2. (c) The positive defocus scenario. (d) The negative defocus scenario.

positive / negative defocus according to the locations where defocus happens with respect to the focal plane.

The PSF of Spherical Aberration In Microscopy

It is necessary to analyze the PSF of spherical aberration in positive / negative defocus scenarios. In the literature, Luo *et al.* [52] measured the PSF of a microscope by creating a 3D PSF z-stack (40×/0.95NA objective lens; 300nm fluorescence polystyrene latex beads; the z-stack from -10μm to 10μm with 0.2μm axial steps). In their 3D PSF model, we find that spherical aberration in positive / negative defocus scenarios produces an asymmetrical PSF. It indicates that diffraction rings with positive defocus and a blur speckle with negative defocus, which is consistent with our observation. Furthermore, experiments in [52] also demonstrate that the asymmetry effect results from the spherical aberration, rather than the thickness of samples (the size of beads is negligible). Besides, we give the out-of-focus degradation imaging model

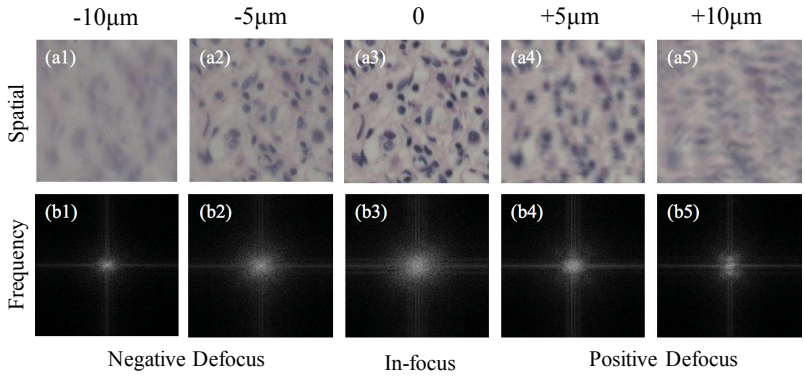

Figure 3.4: The pathology images (a) and the corresponding frequency images (b) at different defocus distances.

in the appendix for readers interested.

3.2.3 Defocus Images in WSI

In this subsection, we study the differences between positive / negative defocus images caused by optical aberrations. We exhibit the pathology images and the corresponding frequency images at the different defocus distances, as shown in Fig. 3.4. We get distinctive statistics from abundant sample images captured by WSI with positive / negative defocus offsets. The negative defocus images have more uniform blur, while the positive ones have visible artifacts, such as stripes. The corresponding frequency images also have noticeable differences: the positive defocus frequency image has a central peak and two secondary peaks. The state-of-the-art [30] did not notice the asymmetry effect of optical aberrations. In [30], the network learns one single mapping function for all images. While we propose a binary classification network to discriminate the sign of defocus offsets, which will be elaborated in the next section.

In conclusion, we can classify positive and negative defocus first according to the features of pathology images. Then, we perform algorithm processing in the same defocus category.

3.3 The Proposed Method

3.3.1 The Proposed Autofocusing Method

To realize autofocusing, the most popular method in current WSI systems is focus map surveying [42], which creates a focus map after tile-by-tile scanning by z-stacks, as shown in Fig. 3.11 (a). However, in our method, the focus predicted process from a z-stack is replaced by a neural network with a single shot, as shown in Fig. 3.11 (b). The input of network is a single defocus tile image, and the output is the predicted defocus distance for the tile. Then, defocus distances congregate together to generate a focus map, by which the microscope scans the sample and performs shooting. Therefore, the difference is that ours only needs to take a single shot, while the traditional method needs to create a z-stack (n times shoots with the corresponding mechanical z-scanning, $n = 21$ [42]). In conclusion, the merits of our methods are high accuracy, high speed and compatibility.

We introduce our WSI workflow based on deep learning autofocusing in detail in the next subsection.

3.3.2 Deep Cascade Networks Overview

In this paper, we leverage the knowledge of physics-based observation along with a neural network architecture [58, 75, 94, 74], and propose a learning-based strategy

for autofocusing via deep cascade networks. Deep cascade networks usually combine multi-stages for multi-tasks to train separately and test jointly, such as cascaded classifiers [85], deep coarse-to-fine cascade networks [70] and reconstructing dynamic sequences and each frame independently [71]. Although a single network may be powerful adequately to learn one step reconstruction, such a one-step network could show signs of overfitting, unless there are sufficient data to train [71]. Besides, a one-step network may require a long time to train and fine-tune carefully.

A simple and effective solution is to train a second network independently, which learns features and signs from the output of the first network. Therefore, we develop a learning-based strategy for autofocusing via deep cascade networks, containing defocusing classification network and refocusing network. As shown in Fig. 3.5-(a), in our cascade networks, the input is a defocus image and the output is the predicted defocus distance. More specifically, the input defocus image is firstly divided into subimages, for each of which the classification network is conduced to discriminate the sign of the defocus offset. Then the accurate defocus distance is identified by the refocusing network for positive / negative offset respectively. Finally, autofocusing in WSI is realized by shifting mobile platform of microscope to the corresponding defocus distance position. We introduce in detail the designs of these main modules in the following.

3.3.3 Defocus Classification Network Design

Based on the asymmetry effect of aberrations, we design a defocus classification network to distinguish sample images with positive or negative defocus offset, as shown in Fig. 3.5-(b). More specifically, the defocus offset is considered negative when the

sample slide is situated outside or at the focal length; otherwise, the defocus offset is positive.

In our network, the input is a defocus image, the output are binary decisions indicating positive / negative defocus. In the network training stage, the binary decisions are derived from the signs of labeled defocus distances. We design a CNN network with channel attention. Specifically, a convolutional layer (5×5, stride 1) extracts low level features from the input defocus images. Subsequently, we utilize a channel attention layer to weight features and a max-polling layer to reduce dimensionality [27]. Then we repeat the block (convolution + channel attention + max-pooling) four times and use a pointwise convolution to fuse the information. Finally, two fully connected layers classify defocus features as positive or negative. The final fully connected layer with a softmax activation function outputs label 1 for a positive sign and 0 for a negative sign.

In network training, we adopt cross entropy as the loss function

$$L_1 = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i)], \tag{3.3.1}$$

where y_i is the predicted sign, $\hat{y}_i$ is the sign of the i-th labeled defocus distance, and n is the number of images in each batch.

3.3.4 Refocusing Network Design

After the defocus classification network is designed, samples are classified into two categories. It is known that samples within the same class share similar characteristics. Therefore, we design two-branch refocusing networks—positive network and negative network—to identify defocus distances for two categories respectively.

We determine two-branch networks with the same structure, because samples are captured by the same optical system, as shown in Fig. 3.5-(c). After the feature extraction of convolution layer (5×5, stride 1) and downsampling of max-pooling, we repeat two similar attention residual blocks (ARB_v1 and ARB_v2) four times. ARB_v1 is used to extract features and ARB_v2 reduce their dimensionality. Finally, the first fully connected layer connect all features and the second one output the predicted defocus distances.

In network training, the loss function is defined as follows:

$$L_2 = \frac{1}{n}\sum_{i=1}^{n}(D_i - \tilde{D}_i)^2, \tag{3.3.2}$$

where D_i is the ground-truth defocus distance and $\tilde{D}_i$ is the predicted one.

We define the focus estimation error $D_{MAE} = |D_i - \tilde{D}_i|$. We can obtain faithful autofocusing results as long as $D_{MAE} < D_{DOF}$. It offers a flexibility for the proposed refocusing network, *i.e.*, the result estimated by the refocusing network is not necessary to be the exact defocus distance, but just within the DoF.

3.3.5 Networks Training

Autofocusing Dataset

In networks training, we utilize the dataset collected by Jiang *et al.* [30], which includes about 130,000 images with the corresponding defocus distances. To be fair, we adopt the same training, validation and test sets as [30]. The training set includes 35 research-grade human pathology slides with Hematoxylin and eosin stains (Omano OMSK-HP50). The images were obtained by a color camera with pixel size $3.45\mu m$.

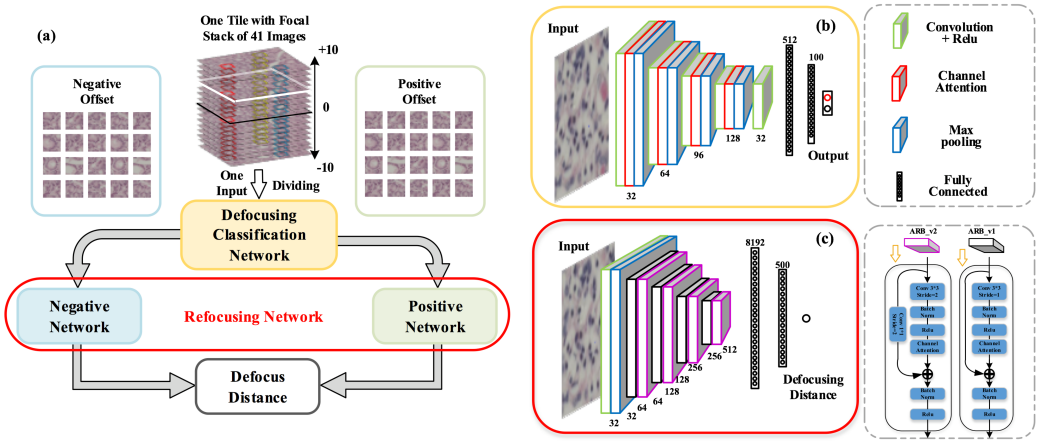

Figure 3.5: An overview of proposed framework. (a) Overall framework of proposed autofocusing cascade networks. The input is a defocus image from focal stack and the output is the predicted defocus distance. (b) The defocus classification network. The output stands for positive or negative label. (c) The refocusing network. Positive and negative networks have the same structure.

They are further divided into 224×224 smaller segments for further usage. A typical WSI system uses a 0.75NA, 20× objective lens to acquire high-resolution images of the sample. We follow the same setting in our network.

In the collection of sample dataset, a sequence of z-stack images are acquired with 41 different defocus distances and step size $0.5\mu m$ from -$10\mu m$ to +$10\mu m$, which are sufficient to cover the possible focus offset. The in-focus ground truth is recovered by maximizing Brenner gradient [90] of the z-stack images and it is considered to be the reference plane. By shifting the axial mechanical stage from the reference plane, the defocus images are obtained and corresponding defocus distances are recorded as labels.

Implementation

The outputs of two sub-networks are defocus offset signs and defocus distances, respectively. We choose to optimize two sub-networks separately, and finally cast them together as a cascade autofocusing network.

In the classification network, we use all defocus images with labeled distances as our training set. Our classifier is trained using the ADAM optimizer with a learning rate as 0.0001 for 50 epochs and with batch size as 128. The training time is about 14 hours. In the refocusing network, we select the positive / negative labeled images to train positive / negative networks. Dropout rate 0.3 is employed for the first fully connected layer to suppress overfitting. The refocusing network is also trained using the ADAM optimizer with a learning rate as 0.0005 for 50 epochs. The batch size is 128 and training time is about 8 hours. All networks training is run on a single NVIDIA GTX 1080Ti.

3.4 Experiments

In this section, we provide performance comparison of defocus distance prediction with the SOTA [30], which is the first learning-based autofocusing method for WSI. Their methods all use a ResNet-50.

We provide experimental comparisons on two sets: 1) **Dataset 1** built by [30] are the same vendor with the training data. It contains all stained tissue slide images, including six categories of biological tissues with different morphological characteristics of size, thickness and structure. Each sample contains 41 images from -10μm to +10μm with interval 0.5μm. 2) **Dataset 2** contains the de-identified HE skin-tissue

Table 3.1: The focusing error comparison of ours and three varaints of [30] under incoherent illumination on Dataset 1.

Sample	[30] in spatial-domain	[30] in Fourier-domain	[30] in dual-domain	Proposed method
Sample1	0.33 ± 0.25	0.61 ± 0.58	**0.27 ± 0.18**	0.29 ± 0.22
Sample2	**0.33 ± 0.26**	0.70 ± 0.83	0.96 ± 0.86	0.62 ± 0.49
Sample3	0.37 ± 0.22	0.53 ± 0.35	**0.31 ± 0.22**	0.33 ± 0.21
Sample4	0.53 ± 0.28	0.50 ± 0.34	0.42 ± 0.24	**0.34 ± 0.25**
Sample5	0.58 ± 0.31	0.63 ± 0.39	**0.36 ± 0.29**	0.39 ± 0.30
Sample6	0.87 ± 0.57	0.70 ± 0.52	0.45 ± 0.24	**0.41 ± 0.26**
Summary	0.50 ± 0.32	0.61 ± 0.50	0.46 ± 0.34	**0.37 ± 0.31**

Table 3.2: The focusing error comparison of ours and three varaints of [30] under incoherent illumination on Dataset 2.

Sample	[30] in spatial-domain	[30] in Fourier-domain	[30] in dual-domain	Proposed method
Sample7	1.51 ± 1.02	0.94 ± 0.71	0.48 ± 0.32	**0.42 ± 0.25**
Sample8	1.32 ± 1.29	0.99 ± 1.51	1.03 ± 1.50	**0.72 ± 1.46**
Sample9	2.69 ± 2.41	0.63 ± 0.50	**0.28 ± 0.28**	0.36 ± 0.29
Sample10	2.19 ± 2.15	0.77 ± 0.53	**0.38 ± 0.38**	0.40 ± 0.30
Sample11	2.19 ± 2.15	0.77 ± 0.53	0.43 ± 0.69	**0.37 ± 0.34**
Sample12	1.00 ± 0.77	**0.52 ± 0.29**	0.85 ± 0.73	0.76 ± 1.76
Sample13	2.19 ± 2.15	0.77 ± 0.53	**0.29 ± 0.22**	0.33 ± 0.53
Summary	1.85 ± 1.68	0.71 ± 0.62	0.53 ± 0.49	**0.46 ± 0.90**

slides made by the Dermatology Department of the UConn Health Center, which are different sources from the training data [30]. For both datasets, the size of each tile image is 2448×2048.

3.4.1 Performance Comparison of Defocus Distance Prediction under Incoherent Illumination

Comparison of Focusing Errors

In WSI, the most widely used objective criterion for performance evaluation of defocus distance prediction is the **focusing error**, which represents the differences of predicted defocus distance with respect to the ground truth D_{GT}. The focusing error

is measured by mean absolute error (MAE) D_{MAE} and standard deviation (SD) D_{SD}.

The objective comparison results of focusing errors on Dataset 1 are exhibited in Tab. 3.1. Tab. 3.1 shows the comparison results under incoherent illumination, where three variants of [30] are compared: 1) spatial-domain-only method, which exploits RGB channels information only; 2) Fourier-domain-only method, which exploits the Fourier domain information with a magnitude channel and a angle channel; 3) dual-domain method, which combines spatial and Fourier domain information. For the sake of fairness, the compared schemes do not involve any hardware modifications. With respect to the average focusing errors over six test samples, our method achieves the best performance compared with all three variants of [30]. We further provide comparison results on Dataset 2. Tab. 3.2 exhibits results under incoherent illumination. Our method still achieves the best prediction performance among all three variants of [30].

We exhibit the subjective autofocusing performance in WSI on Dataset 1, as shown in Fig. 3.6, in which the left images are in-focus specimens as ground truth in each sample. The top right are the out-of-focus images in the defocus distance of -10μm, -5μm, 5μm and 10μm, respectively. The bottom right are the autofocusing performances of the corresponding defocus images. *Samples* from *1* to *6* are de-identified HE skin-tissue slides with significantly different biological structural features. For example, *Sample 4* is with more prominent edge features, and *Sample 3* is with more distinct nuclear structure. In contrast, *Sample 2* contains large transparent regions and weaker structural features. For this case, our performance is worse than [8], since the features in *Sample 2* are not so classifiable and thus the binary classification module does not work well. Experiments have proved that the proposed method

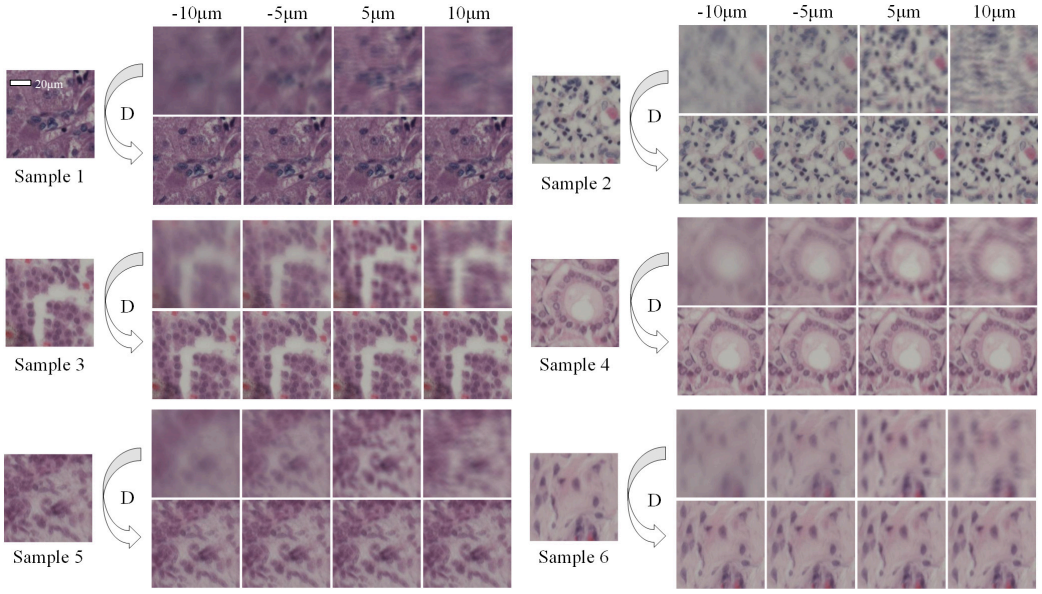

Figure 3.6: The autofocusing performance on Dataset 1. The left images are in-focus specimens as ground truth in each sample. The right images are defocus images and the corresponding focusing performances.

can estimate the focus distance and achieve autofocusing for most of samples with different features. Additional objective evaluation index for WSI is not required in autofocusing performance except for the focusing error.

Comparison of Focusing Errors with respect to DoF

The predicted defocus distances are not necessary to be the exact ones. We can obtain in-focus images as long as the focusing errors are less than DoF of the objective lens. Accordingly, we make the comparison of focusing error to DoF to demonstrate the effectiveness of our autofocusing performance.

We show the average focusing error distribution of our method on Dataset 1 and 2 in Fig. 3.7 under incoherent illumination. Each point stands for the average focusing

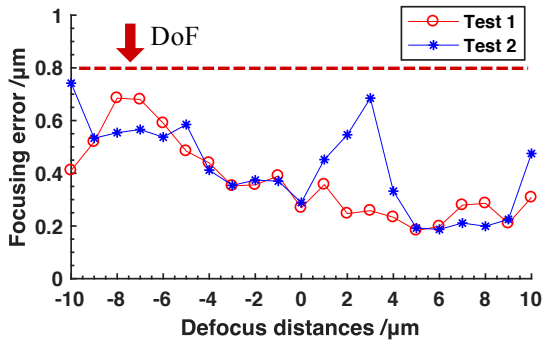

Figure 3.7: The average focusing error distribution on Dataset 1 and 2. Each red or blue point stands for the average focusing error of different defocus distances under incoherent illumination.

error at different defocus distances. In a typical WSI system (0.75NA, 20× objective lens), the DoF is $0.8\mu m$ calculated by Eq. 1 according to the hardware parameters. It can be found that, all average errors are within the range of DoF in Fig. 3.7. In fact, there are 92.25% of focus errors within the range of DoF on Dataset 1 and 89.48% on Dataset 2, with the defocus distance from -10μm to +10μm. The average thickness of the pathological tissue is usually $5\mu m$. Thus, we can focus on the data with defocus distance from -5μm to +5μm, and there are 97.71% of focus errors within the range of DoF on Dataset 1 and 95.12% on Dataset 2. Therefore, the focusing error distribution of our method is more concentrated in the range of DoF. However, there are much more points outside of DoF in dual-domain method [30] than those of ours. This analysis demonstrates the superiority of our method in terms of accuracy.

In addition, in Fig. 3.7 we find that the points of focusing errors have significant differences in spatial distribution. The distribution of positive points is more concentrated within the range of DoF, while negative points have a more dispersed distribution. The methods of [30] also show similar results. These results are the

consequence of the asymmetry effect of optical aberrations: the positive defocus images have more distinct optical artifacts, which contribute to extract features by the network; while the negative ones have more uniform defocus blur, that is difficult to predict the defocus distances accurately by the network. Experiments prove the rationality of our motivation for defocus classification design.

3.4.2 Performance Comparison of Defocus Distance Prediction under Single-LED Illumination

How effective is our method on other optical modification systems? In this subsection, we compare the methods between ours and the Single-LED method of [30], which utilizes a single green channel input under single-LED illumination condition, rather than the typical incoherent Kolner illumination. Although it is not a typical modification in WSI, it does not affect our evaluation of the network.

Fig. 3.8 illustrates the case on Dataset 1 and 2 under single-LED illumination. It can be found that, the average predicted errors of our method can be reduced by 23% on Dataset 1 and 70% on Dataset 2, compared with Single-LED method of [30]. Only for dyed deeply *Sample 1* and dyed slightly *Sample 2*, our performance is worse than [30].

According to the figure, we find that: (1) Under single-LED illumination, most of the focusing errors are less than DoF of the objective lens. However, the focusing errors of Single-LED method in [30] are higher than DoF, even up to 3 times of DoF on *Sample 8*. These results demonstrate that our algorithm is capable of achieving qualified autofocusing. (2) Compared with the scenario under single-LED illumination, there are more focusing errors within the range of DoF under incoherent illumination.

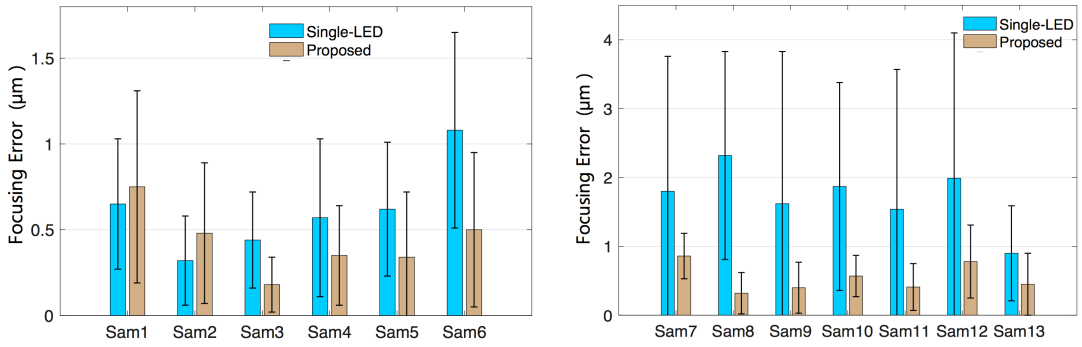

Figure 3.8: The focusing error comparison of ours and Single-LED of [30] on Dataset 1 (left) and Dataset 2 (right) under single-LED illumination.

It is the reason that the three channels of RGB under incoherent illumination contain more characteristics information provided by RGB channels than the single-channel under single-LED illumination. Therefore, the autofocusing has better performance under incoherent illumination.

3.4.3 Comparison with Other Methods

We have provided autofocusing performance comparison of defocus distance prediction with the state-of-the-art [30], which is the first learning-based autofocusing method for WSI. For the sake of fairness, we only compare with the single defocus image methods, without any optical hardware modification and multi-image inputs. How about the performance of other methods? In this subsection, we discuss the schemes involving optical hardware modifications and multi-image inputs.

Table 3.3: The performance of focusing error on Dataset 1. Left: comparisons with [30] under green LED illumination. Right: comparisons with three varaints of [30] under RGB illumination.

Sample	Dual-LED 3-Domain	Proposed method
Sample1	**0.16 ± 0.12**	0.29 ± 0.22
Sample2	**0.26 ± 0.27**	0.62 ± 0.49
Sample3	**0.16 ± 0.11**	0.33 ± 0.21
Sample4	**0.22 ± 0.16**	0.34 ± 0.25
Sample5	**0.17 ± 0.12**	0.39 ± 0.30
Sample6	**0.28 ± 0.22**	0.41 ± 0.26
Summary	**0.21 ± 0.17**	0.37 ± 0.31

Sample	Dual-LED 3-Domain	Proposed method
Sample7	0.80 ± 0.68	**0.42 ± 0.25**
Sample8	**0.52 ± 0.57**	0.72 ± 1.46
Sample9	0.52 ± 0.30	**0.36 ± 0.29**
Sample10	0.73 ± 0.47	**0.40 ± 0.30**
Sample11	0.52 ± 0.39	**0.37 ± 0.34**
Sample12	0.78 ± 0.39	**0.76 ± 1.76**
Sample13	0.33 ± 0.20	**0.33 ± 0.53**
Summary	0.59 ± 0.43	**0.46 ± 0.90**

Comparison with the Hardware Modification Method.

Dual-LED 3-Domain is a hardware modification method, which achieves the best performance in [30]. The 3-domain input under dual-LED illumination contains a spatial intensity channel, a Fourier magnitude channel and an auto-correlation channel. The comparison results on Dataset 1 and 2 are exhibited in Tab. 3.3.

We find that: Although Dual-LED 3-Domain method achieves the better performance on Dataset 1, ours has better performance on Dataset 2. In focusing error distribution on Dataset 2, ours has a more uniform distribution. However, the focusing error distribution of Dual-LED 3-Domain has a steep distribution near the focus position, and the focusing error is even up to $2.5\mu m$ in [30]. Therefore, our method achieves better performance compared with the hardware modification method (Dual-LED 3-Domain), which shows the best performance of [30].

It is worth mentioning that, our method enjoys the merits of compatibility and low costs, because of no modifications on the optical hardware system.

Comparison with the Multi-image Method.

The method of [66] utilizes the difference image of two defocus images as the network input. The interval between the two defocus images is $2\mu m$. All data sets are also from [30]. The average focusing errors of this method are lower than those of our method, with $0.22 \pm 0.25\mu m$ on Dataset 1 and $0.36 \pm 0.37\mu m$ on Dataset 2.

Our method is a single-shot method, which only utilizes one defocus image to predict the defocus distance. The shooting position is fixed along the z-axis, so we only scan the slide along the x-y direction to create a focus map. However, the method of [66] employs two defocus images, which need to re-scan the slide along the z-axis and shot for a second time. Additional scanning and shooting significantly reduce the speed of the WSI workflow. Therefore, our single-shot method is more suitable for WSI.

3.4.4 The Necessity Analysis of Defocusing Classification Network

The main contribution of this work is the binary classification network that exploits the asymmetry effect of optical aberrations. In this subsection, we demonstrate the necessity of classification by experimental analysis on Dataset 1 and 2.

Performance of Defocus Classification Network

Due to the effect of the non-uniformity of sample thickness, we perform the same data pre-processing as [30], which divides the test image into 20 sub-images with 224×224 regions and discards outliers. These sub-images are used as the input of the proposed deep cascade network. We select 20 non-overlapping regions as the basis for the

Table 3.4: The focusing error comparison of positive refocusing network R_p and negative refocusing network R_n on Dataset 1 and 2

Test set	R_p (Incoherent)	R_n (Incoherent)
Dataset 1	**0.25 ± 0.18**	0.48 ± 0.35
Dataset 2	**0.25 ± 0.18**	0.51 ± 0.44

overall consideration. When determining whether a test image is positive or negative defocus, we perform classification on these regions and count how many ones with positive or negative labels. The sign that has the maximum number of regions are considered to be the type of this test image. The accuracy rate of our classification network is 98.85% on Dataset 1 and 97.48% on Dataset 2. The experimental results demonstrate that the defocusing classification network we designed has satisfactory performances.

Performance of Refocusing Network

Refocusing networks contain two parts: positive network R_p and negative network R_n. The comparison focusing errors of refocusing networks on Dataset 1 and Dataset 2 are exhibited in Tab. 3.4. We find that, for two-branch refocusing networks with the same structure, the results of defocus distance prediction are significantly different between positive and negative scenarios. The positive focusing errors are lower than negative ones, about 50.5%.

The results demonstrate that, under the influence of asymmetry optical aberrations, negative defocus images have a more uniform defocus distribution than positive defocus images. The positive images have visible artifacts, which contributes to network identification and classification. In contrast, the negative defocus images with a uniform defocus distribution bring difficulty to network classification.

Table 3.5: The focusing error comparison of four methods on Dataset 1 and 2

Methods	Dataset 1	Dataset 2
Baseline (Incoherent)	0.50 ± 0.32	1.94 ± 1.91
State-of-the-art (Incoherent)	0.46 ± 0.34	0.53 ± 0.59
Refocusing (Incoherent)	0.41 ± 0.33	0.60 ± 0.61
Classification+Refocusing (Incoherent)	**0.37 ± 0.31**	**0.46 ± 0.90**

Table 3.6: The focusing error comparison of positive refocusing network R_p and negative refocusing network R_n on Dataset 1 and 2 under single-LED illumination.

Test set	R_p (LED)	R_n (LED)
Dataset 1	0.49 ± 0.47	0.46 ± 0.47
Dataset 2	0.52 ± 0.42	0.49 ± 0.37

Necessity Classification Analysis

The necessity classification analysis is performed on four scenarios in Tab. 3.5: 1) *Baseline (Incoherent)*: a ResNet-50 network, which is the approach in [30], to predict defocus distances directly without classifcation; 2) *State-of-the-art (Incoherent)*: a ResNet-50 network, which is the dual-domain approach in [30]; 3) *Refocusing without Classification (Incoherent)*: refocusing network with all defocus images trained together without classification; 4) *Classification + Refocusing (Incoherent)*: our deep cascade networks, including the classification network and the refocusing network. In this comparison study, we can investigate the role of the binary classification network fairly.

As indicated in Tab. 3.5, the performance of our deep cascade networks is remarkably better than refocusing network without classification. The average predicted defocus distance errors can be reduced by 9.76% on dataset 1 and 23.33% on dataset 2. From this analysis, it can be found that the classification before refocusing is necessary, and our proposed strategy is effective.

Table 3.7: The focusing error comparison of four methods on Dataset 1 and 2 under the single green LED illuminaiton

Methods	Dataset 1	Dataset 2
Baseline (LED)	0.61 ± 0.39	1.72 ± 1.72
Refocusing (LED)	0.51 ± 0.46	1.70 ± 1.33
Classification+Refocusing (LED)	**0.47 ± 0.42**	**0.51 ± 0.42**

Analysis of Classification under Single LED Illumination

The main contribution of this work is the binary classification network that exploits the asymmetry effect of optical aberrations. We demonstrate the necessity of classification by experimental analysis on Dataset 1 and 2 under single-LED illumination. The accuracy rate of the defocusing classification network is 98.85% on Dataset 1 and 97.48% on Dataset 2. The comparison focusing errors of refocusing networks on Dataset 1 and 2 are exhibited in Tab. 3.6. We find that, for two-branch refocusing networks with the same structure, the results of defocus distance prediction are similar between positive and negative scenarios.

The necessity classification analysis is performed on three scenarios in Tab. 3.7: 1) *Baseline (LED)*: a ResNet-50 network, which is the approach Single-LED in [30], to predict defocus distances directly without classifcation; 2) *Refocusing without Classification (LED)*: refocusing network with all defocus images trained together without classification; 3) *Classification + Refocusing (LED)*: our deep cascade networks, including the classification network and the refocusing network. In this study, we can investigate the role of the binary classification network objectively. As indicated in Tab. 3.7, the performance of our deep cascade networks is remarkably better than a refocusing network without classification. The average predicted defocus distance

Table 3.8: The classification accuracy comparison of four ablation ways and ResNet-50 on Dataset 1 and 2

Methods	Test set 1	Test set 2
Baseline	97.85%	85.80%
Baseline+Pre-processing	97.99%	91.46%
Baseline+Pre-processing+Augmentation	96.84%	93.75%
Baseline+Pre-processing+Augmentation+Attention	**98.85%**	**97.48%**
ResNet-50	98.28%	90.96%

errors can be reduced by 7.84% on Dataset 1 and 70% on Dataset 2. From this analysis, it can be found that the classification before refocusing is necessary, and our proposed strategy is useable in other optical modification systems.

3.4.5 Ablation Study

For the sake of high accuracy of defocus distance prediction, we use data pre-processing and augmentation methods. Specifically, for the raw defocus data, we use a channel normalization to enhance contrast and highlight features. Then to suppress overfitting, we utilize the color channel data augmentation [76], which transforms RGB to GBR or other color orders. The augmented data also reduces the color sensibility, resulting from histological staining. After balancing data capacity and training time, we add two color orders (GRB &GBR) with distinct color features. Besides, channel attention is a practical approach to weight features.

Defocusing Classification Network

The ablation analysis of classification network is performed on four conditions in Tab. 3.8: a) *Baseline*: a classification network without any channel attention layers and data pre-processing. The accuracy of classification is 97.85% on Dataset

Table 3.9: The focusing error comparison of four ablation ways and ResNet-50 on Dataset 1 and 2

Methods	Test set 1	Test set 2
Baseline	0.27 ± 0.23	0.35 ± 0.28
Baseline+Pre-processing	0.27 ± 0.22	0.30 ± 0.22
Baseline+Pre-processing+Augmentation	0.26 ± 0.25	0.28 ± 0.27
Baseline+Pre-processing+Augmentation+Attention	0.25 ± 0.18	**0.25 ± 0.18**
ResNet-50	**0.21 ± 0.20**	0.33 ± 0.27

1, while the generalization ability of the network is relatively weak on Dataset 2. b) *Baseline+Pre-processing*: a classification network with data pre-processing. The performance is higher than the last one on all dataset, especially on Dataset 2. c) *Baseline+Pre-processing+Augmentation*: a classification network with data pre-processing and augmentation. The accuracy of classification still increases about 2% on Dataset 2, although accuracy decreases a little on Dataset 1. d) *Baseline+Pre-processing+Augmentation+Attention*: our defocusing classification network, including data pre-processing, augmentation and channel attentions. In this scenario, we get the highest performance on both Dataset 1 and 2. Therefore, in a classification network ablation study, the application of data pre-processing, augmentation and channel attentions indicate their effectiveness and practicability.

Besides, we take a typical ResNet-50 for the objective comparison. The accuracy of our classification network is slightly higher than ResNet-50 on Dataset 1, while our performance on Dataset 2 is 97.48%, which is much higher than ResNet-50 90.96%. Besides, the parameters of our classification network are 8MB, while the parameters of ResNet-50 are up to 270MB. In summary, our defocusing classification network enjoys the following merits: only a few parameters, high speed and strong generalization ability.

Table 3.10: The comprehensive comparison between ResNet-50 and ours.

Method	ResNet-50	Ours
Focusing Errors (μm)	0.50 (Dataset 1)	**0.37** (Dataset 1)
	1.85 (Dataset 2)	**0.46** (Dataset 2)
Parameters (MB)	270	**248**
Inference Time (s)	**89.6**	90.4

Refocusing Network

The ablation analysis of refocusing network is performed on four conditions in Tab. 3.9, which are the same as the classification network. Likewise, in refocusing network ablation study, the application of data pre-processing, augmentation and channel attentions demonstrate their effectiveness of defocus distance prediction. We also utilize a ResNet-50 for the objective comparison of focusing error. All networks are trained and tested by positive defocus images only. Although the focusing error of our refocusing network is slightly lower 16% than ResNet-50 on Dataset 1, our performance is higher 24.24% than ResNet-50 on Dataset 2. In addition, the parameters of our refocusing network are 120MB, while the parameters of a ResNet-50 are up to 270MB. Therefore, the refocusing network has more advantages than typical ResNet-50.

The Comprehensive Comparison between ResNet-50 and Ours

The inference time of ResNet-50 and ours is $89.6s$ and $90.4s$ on both datasets, respectively. Although our parameters are lower 9% than ResNet-50, ours still has the same inference time as ResNet-50, due to the additional CPU cost. We compare the performance, parameters, and time between ResNet-50 and ours as shown in Tab.3.10. The experiment demonstrates that our cascaded networks (defocusing classification + refocusing) can significantly reduce focusing errors.

3.4.6 The Influence Analysis about the Count of Network Parameters

In this subsection, we will prove that the autofocusing performance improvement results from classification design, rather than the network with more parameters or deeper layers. For a further fair comparison, we design a new deeper single refocusing network with 64 layers into a comparison study. We also design deep cascade networks with 28 layers for the classification module and 35 layers for the refocusing module. All of them use the ResNet, which is the same network as [30], except the number of network layers. In this way, we can see more clearly the role of the proposed binary classification network to the final performance. The comparison group now includes three cases:

- The single regression network with 54 layers (ResNet) proposed by [30], which is the baseline.
- The deeper single refocusing network with 64 layers (ResNet).
- The proposed cascade network with 28 layers (ResNet) for classification module and 35 layers (ResNet) for refocusing module.

First, let us investigate the performance comparison between the baseline and the deeper baseline. As shown in Fig. 3.9, for six test samples, the deeper baseline network achieves better performance on *Sample 3, 4* and *6*, but loses in the rest ones. So the comparison result is a tie. It means that simply increasing the network layers is not a straightforward and inevitable manner to improve the performance of autofocusing. Furthermore, we check the performance comparison between the deeper baseline and ours, which are with almost the same network parameters. From

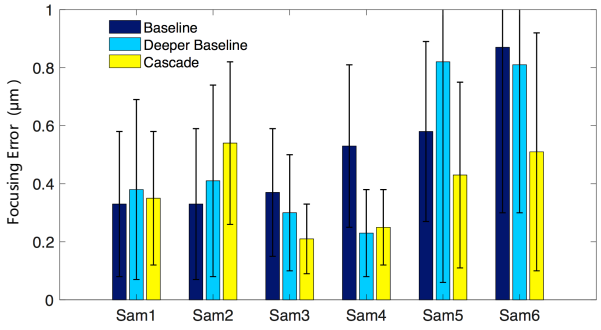

Figure 3.9: The ablation analysis about the count of network parameters. Baseline: the method from [30] with 54 layers; Deeper Baseline: deeper refocusing network with 64 layers; Cascade: cascaded networks with classification network (28 layers) and refocusing network (35 layers).

Fig. 3.9, it can be found that our scheme wins on *Sample 1, 3, 5* and *6*, and slightly loses on *Sample 4*. The average defocus distance error of the deeper baseline is $0.43\mu m$; in contrast, that of ours is $0.36\mu m$. The above analysis demonstrates that the proposed binary classification module is constructive to improve the performance of autofocusing.

In general, pathology slides can be categorized as small tissue sections (mouse testis or TMA cores), medium tissue sections (mouse brains), and large tissue sections (animal embryos). Our training images are human H&E stained pathology slides with uniform thickness $4\sim5\mu m$, mainly specific to the small tissue sections. The defocus distance is from -10 to $+10\mu m$, *i.e.*, $20\mu m$, which is sufficient for our task. (For the thick and large tissue sections, we need to add more defocus images with longer defocus distances to the training set.)

It is a general question if the defocus distance exceeds the training dataset range. We just need to expand the range of the training samples and add these samples to

the training set, if we want to adjust the model trained with small range samples to defocus samples with large range samples. Here, we design an experiment to demonstrate our point. We take $0\sim+5\mu m$ dataset for training and $0\sim+10\mu m$ dataset for testing:

- To be objective, we utilize a typical network: Resnet-50;
- We train Resnet-50 by two datasets respectively: the first network is trained by the data with defocus distances $0\sim+5\mu m$ (0-5 ResNet) and the second network is trained by the images with defocus distances $0\sim+10\mu m$ (0-10 ResNet);
- The testing set contains the images with defocus distances $0\sim+10\mu m$ for two networks;
- The two networks have the same hyperparameters;
- We only show the performance on the positive dataset. (The negative has a similar performance.)

The comparison of errors at different defocus distances exhibits in Fig. 3.10. We find that: (1) at $0\sim4\mu m$, the 0-5 ResNet achieves similar performance to the 0-10 ResNet; (2) at $5\mu m$, the 0-5 ResNet error is double that of the 0-10 ResNet; (3) at $5\sim10\mu m$, the 0-5 ResNet has a much worse performance totally than the 0-10 ResNet. As the defocus distance increases out of the training set range, the error of 0-5 ResNet increases approximately linearly. The experiment demonstrates that, due to the small range of training set, 0-5 ResNet can extract small defocus features but not extract large defocus features.

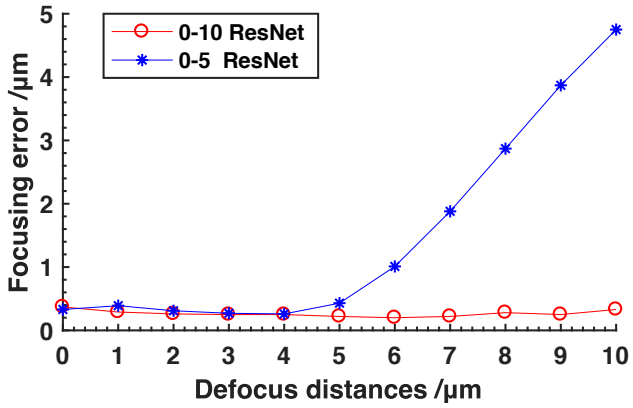

Figure 3.10: The focusing error comparison between 0-10 ResNet and 0-5 ResNet. The ResNet-50s are trained by two datasets ($0\sim+5\mu m$ and $0\sim+10\mu m$), respectively.

In conclusion, although it inevitably increases errors when the network processes the images beyond the training set range, we can adjust the model trained with small range samples to process large range samples by adding large range samples to the training set.

3.5 Applications

In current WSI practices, the most prevalent method is focus map surveying [7], which creates a focus map after scanning each tile through a z-stack, as illustrated in Fig.3.11(a). However, in the approach proposed in this chapter, the process of predicting focus from the z-stack is replaced by a neural network, as depicted in Fig.3.11(b). The network takes a single defocus tile as input and outputs the predicted defocus distance for that tile. Subsequently, all the defocus distances are aggregated to generate a focus map. Using this map, the pathology scanning system scans the

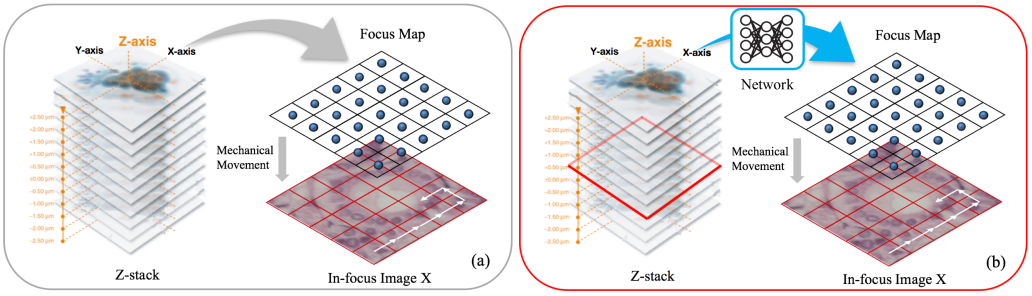

Figure 3.11: The conventional focus map surveying method (a) and our deep neural network autofocusing scheme (b). The input of our network is only a single defocus tile image.

sample and carries out the shot capture. Therefore, the distinction lies in the fact that our method necessitates only a single shot capture, whereas the traditional approach demands the creation of a z-stack (entailing n mechanical z-axis movements and camera exposures, where $n = 21$). The detailed WSI workflow is shown in Appendix A.3.1.

In the conventional z-stack technique, constructing a z-stack entails n axial movements (with n typically being 11, though selecting a higher number can yield better results). The Brenner gradient method is utilized to identify the in-focus plane, necessitating n calculations, with each gradient computation of an image requiring approximately 1.4 s. Assuming each axial movement takes time P, the total time for a single autofocusing operation using the traditional z-stack method is calculated as (10×P + 11×1.4) s. In contrast, the aberration-guided WSI autofocusing approach introduced in this study leverages network inference, completing in a mere 2.5 s. This method necessitates only one axial movement, thereby eliminating the need for additional movements and gradient calculations. Consequently, the time required for our

method to perform a single autofocusing operation is merely (2.5 + P) s. Our findings demonstrate that, in comparison to the traditional z-stack method, our approach markedly decreases the time required for autofocusing.

This method provides several benefits, including minimal in-focus error, swift focusing capability, and robust compatibility, making it particularly suitable for applications that demand accurate exposure captures of pathology images. Nonetheless, there are some limitations to consider. The method exhibits challenges in classifying transparent samples, reflecting limitations in its generalizability. Furthermore, the reliance on network inference necessitates the computational resources of high-performance GPUs.

3.6 Conclusion

This chapter introduces a WSI autofocusing method based on a deep cascading network. Leveraging the asymmetric properties of optical aberrations, a defocus classification network is designed, categorizing samples with distinct feature characteristics into two classes. Benefitting from the classification results, the subsequent two refocusing network branches can effectively learn the mapping between defocus images and defocus distances. This approach can overcome the limitations of traditional methods, facilitating rapid and precise focus prediction, and is compatible with current WSI methodologies. Experimental results indicate that, compared to SOTA focus estimation techniques, this method yields lower focusing errors and is particularly suitable for real-exposure photography of pathological images.

Chapter 4

Dual-shot Deep Autofocusing with a Fixed Offset Prior

4.1 Introduction

WSI is an emerging technology in digital pathology. The accuracy and speed of autofocusing are crucial for the performance of the WSI system. Traditional autofocusing methods require capturing a stack of up to 21 shoots with varying focal distances for each tile of the target ultra-high-resolution pathology image. Conventional autofocusing methods either are time-consuming or require additional hardware and thus are not compatible with the current WSI systems, as mentioned in Chapter 1.1. Compared to mechanically adjusting the focal distance on a tile-by-tile basis, using advanced machine learning algorithms to deblur defocus pathological images is an efficient approach. This method can produce sharp slide images in a single pass without the need to create a focus map or employ expensive and complex optical hardware. However, achieving blind deblurring of pathological images presents challenges under

constraints such as high NA and magnification by objective lenses, including uneven focus distribution and limited DoF.

Diverging from the focus prediction approach outlined in Chapter 3, this chapter introduces a method that is capable of directly restoring images to their in-focus state. To overcome the imaging bottleneck, we have developed a deep convolutional neural network for tile-wise autofocusing, designed to generate in-focus images from tentatively defocus ones. This dual-shot autofocusing network (DAFNet) operates with just two images taken at different focal distances, using their relatively fixed offset as an implicit prior. Through a constrained position design, we utilize two defocus images taken at fixed relative positions to derive a univariate equation for the in-focus image, thereby transforming the problem of blind deblurring into a non-blind deblurring issue. The innovative architecture of DAFNet facilitates the fusion of complementary information from the two input images taken at different focal lengths. The proposed offline reconstruction strategy allows for fast scanning of sample slides without compromising on image quality, as DAFNet is capable of correcting errors in the focal distance and bringing the scanned tiles back into focus through a learned non-linear, dual-input blur-to-sharp mapping. Experimental results showcase the refocusing capabilities of the DAFNet method.

4.2 Preliminaries and Motivations

Traditional autofocusing methods rely on mechanical adjustment to conduct refocusing, which need repetitive axial scanning and thus are time-consuming. In order to reduce the time cost of scanning, we propose the concept of *deep autofocusing*, which no longer performs mechanical autofocusing but instead recovers in-focus images in

a learning-based manner. In this section, we introduce the problem formulation and the motivation of the proposed scheme.

4.2.1 Defocus Degradation Model

In optical microscopy, the PSF can be formulated by the classical Born & Wolf model [8, 26]:

$$h(r, \Delta_D) = \left| C \int_0^1 J_0 \left(k \frac{\mathrm{NA}}{n} r\rho \right) e^{-\frac{1}{2} i k \rho^2 \Delta_D \left(\frac{\mathrm{NA}}{n}\right)^2} \rho d\rho \right|^2, \tag{4.2.1}$$

where

– r is the radial distance along the lateral plane;

– Δ_D is the distance between the in-focus position and the object plane along the optical axis, *i.e.*, the defocus distance;

– C is a normalization constant;

– J_0 is zero-order Bessel function of the first kind;

– k is angular wave number of the light source;

– n is the refractive index;

– i is the imaginary number;

– ρ is the normalized coordinate in the exit pupil.

As shown in the above formulation, the blurring artifact in digital pathology is mainly due to the poor focusing effect induced by ΔD. The axial PSF model is shown in Fig. 4.1 (a) and the lateral planes with different ΔD are shown in Fig. 4.1 (b) and (c). It can be found that, the amplitude of blue line ($\Delta D = 0.5\mu$m) is lower than the red one ($\Delta D = 0$) due to the out-of-focus degradation, which becomes larger as ΔD increases.

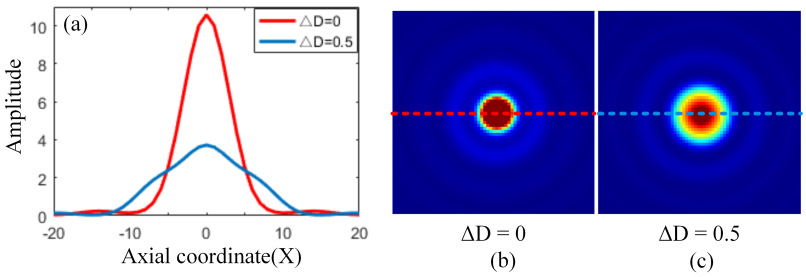

Figure 4.1: (a) The axial PSF distribution curve with in-focus position (red line) and defocus position (blue line). (b) The lateral plane with $\Delta D = 0$. (c) The lateral plane with $\Delta D = 0.5$ μm.

In WSI, the refocusing task is even harder than deblurring in natural image processing, wherein the scene is considered with a constant depth from the camera and thus the PSF is uniform over the image. In contrast, in WSI the tissues have the diversity of thickness, which causes the discontinuity of depths. It is thus impossible to make a perfectly focused image from a single surface, since the corresponding PSF varies spatially. To simulate the DoF effect, we exploit the layered DoF model [73, 86], which converts continuous depth map to approximated discrete depth layers (object planes). Accordingly, the PSF $h(r, \Delta D)$ is rewritten as h_m, where m stands for the position of each depth layer and h_0 is the PSF of the in-focus depth. Each depth layer is blurred by its corresponding PSF with a convolution operation and the blurred depth layers are integrated to form the captured image. Therefore, the in-focus imaging model of WSI can be formulated as:

$$X = \sum_m x_m \otimes h_m, \tag{4.2.2}$$

where x_m is the discrete depth layer of sample with depth m, $\otimes$ is the convolution

operator, X is the underlying in-focus image of the in-focus object plane x_0. According to the formula (2), we find that the in-focus image X is essentially the result of 3D PSF accumulation on the 3D object. Therefore, the in-focus image in digital pathology and the clear image in natural image processing are different with respect to the imaging principle.

When the sample is shifted by offset ΔD from x_0, the new in-focus object plane is denoted as $x_{\Delta D}$ and the corresponding *m-th* depth layer becomes $x_{m+\Delta D}$. The captured defocus degradation image Y can be represented as:

$$Y = \sum_{m} x_{m+\Delta D} \otimes h_m. \tag{4.2.3}$$

This out-of-focus degradation imaging model indicates that the recovery of in-focus image X from defocus image Y is far more challenging than deblurring in natural image processing. The manner that relies on a single defocus image for image recovery—as done by image deblurring—cannot produce satisfactory image quality. Intuitively, to address this ill-posed problem, multiple observed images should be used in order to exploit the complementary information among them. In this work, we utilize two defocus inputs and achieve wonderful refocusing performance. In addition, to recover the in-focus image, it is reasonable to assume that the most reliable knowledge is from the two nearest defocus planes of the in-focus plane [55, 3], denoted as Y_1 and Y_2 respectively:

$$Y_1 = \sum_{m} x_{m+\Delta D_1} \otimes h_m, \quad Y_2 = \sum_{m} x_{m-\Delta D_2} \otimes h_m. \tag{4.2.4}$$

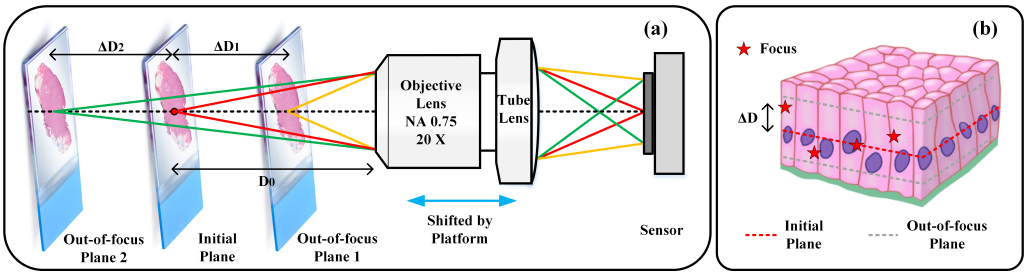

Figure 4.2: Illustration of the microscopy imaging model in WSI system. (a) The proposed dual-shot deep autofocusing scheme. The expected focal distance D_0 (initial plane) over all tiles of the scanned slide, is estimated by performing simple tile autofocusing once from the center filed of the whole slide. Then for all tiles, two tentative possibly defocused images are captured with relative defocus offset ΔD_1 and ΔD_2 to D_0 respectively. (b) Tissue details with many tiles. The focus points of all tiles are different in the range of tissue thickness along the optical axis with an uneven distribution.

The dual-shot defocus images Y_1 and Y_2 retain pieces of complementary information about the underlying in-focus image X, which inspires us to fuse them to obtain the refocused image in a data-driven manner.

4.2.2 Implicit Fixed Offset Prior

Both blind and non-blind deconvolution are techniques in image processing used to restore blurred images, but they differ in terms of problem formulation and solution approaches [35, 20]. Blind deconvolution is a method of deconvolution performed without prior knowledge of the PSF or the blurring kernel. In blind deconvolution, only the blurred image is available, and the task is to estimate both the original image and the blurring kernel. Given the lack of prior information, blind deconvolution presents a more challenging problem. To address blind deconvolution, regularization methods or statistical learning approaches are often employed to constrain the solution space and enhance stability. On the other hand, non-blind deconvolution is

a method where deconvolution is carried out with knowledge of the blurring kernel PSF. In non-blind deconvolution, both the blurred image and the blurring kernel are available, and the task is to directly recover the original image. Compared to blind deconvolution, non-blind deconvolution is typically easier to address because information about the blurring process provides better constraints on the solution space.

We have discovered that utilizing a pair of defocus images for deblurring facilitates the transformation of the challenge from blind to non-blind deblurring. The derivation of the imaging equation for this dual-shot approach yields Eq. A.2.6 (see Appendix A.2). This equation involves merely two unknowns: the defocus distance, denoted as Δ_{D_1}, for the first image, and the relative distance, Δ, between the first and second images. Given the relative distance Δ, Eq. A.2.6 simplifies into a univariate function in terms of Δ_{D_1}. This simplification permits the derivation of an approximate solution via optimization techniques. Following this, the PSF can be explicitly defined by Δ_{D_1}, thereby facilitating the attainment of an in-focus image through the application of non-blind deblurring methods.

In summary, the relative distance Δ serves as an implicit fixed offset prior. Despite the PSF being unknown, the known and predetermined relative distance Δ between the two defocus images allows for the indirect acquisition of the PSF. Consequently, when designing the dual-shot positions, it's crucial to maintain their fixed offset prior. This implicit distance prior does not need to be explicitly specified in the network design. Moreover, while the univariate equation for the in-focus image is known, considering practical factors such as noise and imaging errors, it's not necessary to solve it directly. The recovery of an in-focus image can be achieved through the

implicit input of a neural network.

4.2.3 Defocus Images Determination

In view of the above, we propose a CNN-based autofocusing strategy relying on dual-shot position-constrained images, which are from the two nearest defocus planes of the initial plane, as illustrated in Fig. 4.2 (a). Specifically, at the beginning, we choose the central tile of the scanned slide as the representative one, for which we collect a z-stack with dense images. According to the derived defocus distance, the focal position D_0 is computed, which serves as the initial plane for the subsequent processing. It is worth noting that, since different tiles are with uneven topography, this position is usually not the focus of other tiles. Then for all tiles, two tentative possibly defocus images are captured with relative defocus offset ΔD_1 and ΔD_2 to D_0 respectively. This setup stems from an implicit position prior, namely the relative distance $\Delta = \Delta D_1 + \Delta D_2$. Although we have not explicitly fed the position into the network, the two implicit inputs of the network contain priors of the position. As illustrated in Fig. 4.2 (b), the red star stands for the focus point of each tile and ΔD is the defocus distance from different focus points to the initial plane. The following task is to recover X by fusing its two observations Y_1 and Y_2. This is done by the proposed deep autofocusing network, which will be elaborated in the next section. In practical implementation, we set $\Delta D_1 = \Delta D_2$ for simplicity.

4.3 The Proposed Method

With the dual captured images, we then try to recover the in-focus image with the help of large amounts of training data and high-performance computing environment. In the following, we will introduce the architecture and the training process of the proposed deep autofocusing network (DAFNet) in detail.

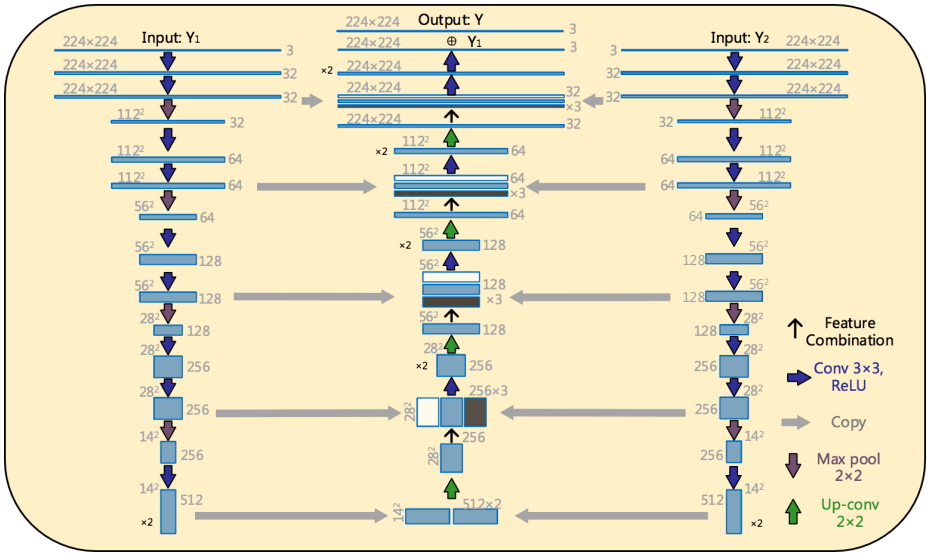

Figure 4.3: The architecture of the proposed DAFNet. Each blue box corresponds to a multi-channel feature map. The number of channels is denoted at the side edge of the box. The x-y-size is provided at the top edge of the box. White boxes represent copied feature maps of the left contracting path. Black boxes represent copied feature maps of the right contracting path. The colorful arrows denote the different operations. $\times$ 2 stands for an additional convolution.

4.3.1 Network Architecture

The network architecture of proposed DAFNet is illustrated in Fig. 4.3. Specifically, the DAFNet consists of two contracting paths (left and right sides) that are with two out-of-focus images Y_1 and Y_2 as inputs, and an expansive path (middle side) that

outputs the recovered in-focus image X. The sharper image of two captured ones is chosen as Y_1, according to the metric of Brenner gradient [79] [90].

- **Contracting paths design:** The contracting paths employ the typical convolutional architecture, including the repeated use of two 3×3 convolutions followed by a rectied linear unit (ReLU) and 2×2 max pooling downsampling layer with stride 2. We double the number of feature channels at every downsampling step. These two paths share the same parameters. Finally, we combine the deepest layers of two paths into a cascaded one.

- **Expansive path design:** The expansive path in each step includes an upsampling feature layer followed by 2×2 convolution (up-convolution), which halves the number of feature channels. We build a concatenation with the corresponding feature maps from the left contracting path (white layer) and the right contracting path (black layer), and employ two 3×3 convolution followed by ReLU. At the final residual layer, Y_1 is added to generate the recovered in-focus image X. In total, the network has 27 convolutional layers.

4.3.2 Network Training

Training Dataset

We use a part of the dataset collected by Jiang *et al.* [30] to train our network. The dataset includes 35 research-grade human pathology slides with Hematoxylin and eosin stains (Omano OMSK-HP50), and contains 162 pathological tissue z-stack tiles. For each tile there is a stack of 41 images taken with different focal distances in a step size of $0.5\mu m$, ranging from -$10\mu m$ to $10\mu m$, with $0\mu m$ corresponding to the

image in focus. The in-focus image is recovered by maximizing Brenner gradient of the z-stack images.

In image stacks of all tiles, the focal distance of an out-of-focus image is given as the defocus offset to the image in focus. But in our system, the microscope camera makes dual shots of each title at two prefixed focal distances. Therefore, we need the out-of-focus images of absolute focal distances to train our DAFNet. We convert the training images in relative focal distance in the dataset of [30] to those in absolute focal distance by simply adding a Gaussian random variable $n \sim \mathcal{N}(0, 1)$ to the relative focal distance. This is because, according to the observation of [24], the focal positions follow a Gaussian distribution, as shown in Fig. 4.4 (a). Specifically, The images of slides are divided into 224×224 patches in Fig. 4.4 (b). Then, we convert the dataset to discrete patches of Gaussian distribution. There are 3240 patches in the initial dataset and we enlarge the dataset by rotation.

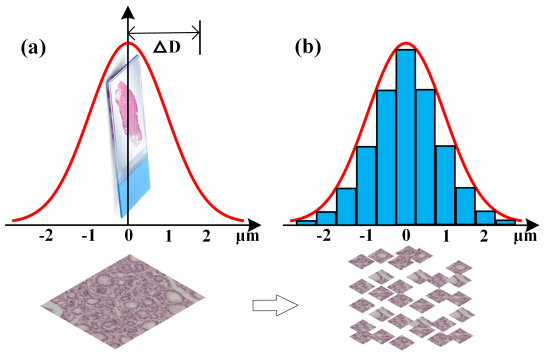

Figure 4.4: Illustration of Gaussian distribution of focal positions. (a) The Gaussian distribution of focal positions with ΔD. The bottom tile shows the continuous fluctuations in the surface of the sample. (b) The discrete Gaussian distribution of focal positions with ΔD. The bottom patches segmented from tiles exhibit the discrete offset of the sample.

Implementation Details

Here we clarify some details in implementation. In network training, the loss function is defined as follows:

$$L = \frac{1}{N}\sum_{i=1}^{N}(X_i - \tilde{X}_i)^2, \tag{4.3.1}$$

where X_i is the ground-truth in-focus image and $\tilde{X}_i$ is the network output, and N is the number of training images in each batch. We select 85% patches with labeled relative defocus offset ΔD as our training set and 15% patches for verification. We utilize batch normalization with batch size as 20 for acceleration training. The network is trained using the ADAM optimizer with a learning rate as 0.0005 for 50 epochs. The network training is run on a single NVIDIA GTX 1080Ti.

4.4 Experiments

In this section, we provide extensive experimental results to demonstrate the effectiveness of our proposed DAFNet scheme.

The experimental analysis is conducted on two public test datasets:

- **Dataset 1:** We use the part of Dataset 1 [30] except that for training as the test set. It contains all stained tissue slide images, including six categories of biological tissues with different morphological characteristics of size, thickness and structure, named *Sample*1 to *Sample*6.

- **Dataset 2:** Dataset 2 [30] that contains the de-identified HE skin-tissue slides made by the Dermatology Department of the UConn Health Center is also used for testing, which is collected from different source with the training set. It

includes seven categories of biological tissues named $Sample7$ to $Sample13$.

For both datasets, the size of each tile image is 2448×2048. We select test images with the corresponding relative defocus offset ΔD ranging from -3μm to +3μm with interval 0.5μm, which are also converted in the same way as the training data. There are 340 and 640 patches in Dataset 1 and Dataset 2 respectively.

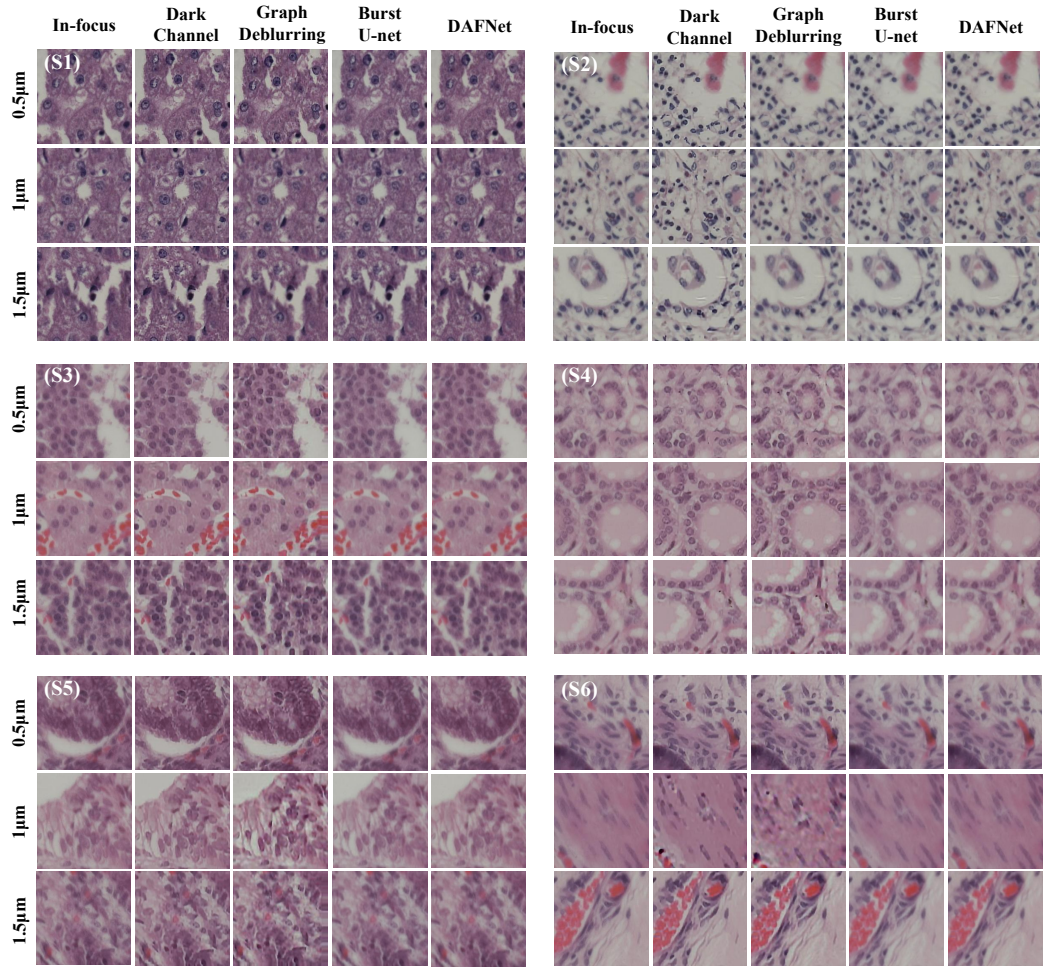

Figure 4.5: Subjective performance comparison on *Sample*1 to *Sample*6.

4.4.1 Comparison with SOTA

In this subsection, to demonstrate the effectiveness of the proposed in-focus image recovery scheme, we provide objective and subjective quality comparison on Dataset 1 with SOTA image deblurring methods, including dark channel prior based [60], graph-prior based [5], U-net based burst deblurring [4] that also takes multiple images as inputs.

The objective performance evaluation with respect to PSNR is shown in Table 4.1, where "-" represents there is no corresponding image in this defocus distance. It can be found that, our method achieves the best PSNR performance on all sample images. These comparison results demonstrate the superior performance of our proposed DAFNet network.

The subjective comparison results on six test images are illustrated in Fig. 4.5. The GT in-focus images are also offered as the quality reference. From the results, it can be found that the statistical prior-based methods, *i.e.*, [60] and [5], cannot handle complicated defocus effects in WSI, since the statistics of biomedical images is different from natural images. These two methods cannot preserve texture information well. Burst U-net based method [4], which also uses deep neural network for burst deblurring, achieves better subjective performance than [60] and [5]. Our method achieves the best subjective performance among compared methods. The recovered in-focus images share a very close subjective effect with the GT in-focus images.

Table 4.1: Objective Performance Comparison with respect to PSNR (dB) of four compared methods.

Methods	ΔD	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Average
Dark [53]	0.5	31.67 ±1.69	29.92 ±0.80	33.89 ±0.88	33.56 ±4.66	32.19 ±3.10	30.30 ±4.69	32.42 ±3.85
Graph [5]		30.57 ±15.59	**51.76 ±0.50**	17.27 ±3.87	22.85 ±8.24	21.09 ±3.85	22.97 ±3.37	26.67 ±12.79
Burst [4]		39.48 ±1.26	38.83 ±0.84	40.62 ±0.26	39.44 ±0.56	39.81 ±0.94	40.41 ±1.22	39.60 ±0.95
DAFNet		**44.83 ±1.48**	48.04 ±0.40	**48.35 ±0.30**	**49.29 ±1.00**	**49.61 ±0.31**	**50.06 ±0.54**	**48.71 ±1.64**
Dark [53]	1	30.74 ±0.90	28.50 ±0.05	32.76 ±0.30	33.37 ±1.16	30.95 ±4.46	30.35 ±2.77	31.74 ±2.77
Graph [5]		28.91 ±9.04	38.79 ±0.33	17.98 ±1.50	19.03 ±3.42	17.38 ±5.09	20.34 ±4.09	21.78 ±7.74
Burst [4]		36.37 ±0.17	35.56 ±0.03	38.03 ±0.04	37.28 ±0.26	37.70 ±0.54	38.50 ±0.81	37.31 ±0.89
DAFNet		**37.13 ±0.34**	**39.16 ±0.34**	**39.81 ±0.24**	**40.37 ±0.36**	**40.44 ±0.16**	**41.35 ±0.78**	**39.93 ±1.34**
Dark [53]	1.5	30.35 ±0.89	28.15 ±1.28	33.65 ±0	32.99 ±0.67	32.43 ±0	31.88 ±0.84	31.35 ±1.73
Graph [5]		20.77 ±8.46	37.50 ±0.65	10.62 ±	20.55 ±1.33	19.17 ±0	23.82 ±4.49	22.19 ±7.78
Burst [4]		34.27 ±0.58	36.84 ±0.20	37.32 ±0	36.80 ±0.47	37.49 ±0	36.92 ±0.58	35.88 ±1.40
DAFNet		**35.18 ±0.48**	**38.05 ±0.56**	**38.91 ±0**	**39.09 ±0.29**	**39.19 ±0**	**39.77 ±0.44**	**37.58 ±2.02**
Dark [53]	2	28.15 ±0.52	28.06 ±0.54	31.06 ±0	32.31 ±0.71	32.28 ±0	30.92 ±0	29.94 ±1.95
Graph [5]		31.61 ±0.67	36.03 ±0.26	19.98 ±0	18.77 ±0.12	26.79 ±0	24.10 ±0	27.53 ±6.47
Burst [4]		32.57 ±0.20	35.23 ±1.05	36.11 ±0	36.10 ±0.27	36.34 ±0	36.86 ±0	34.97 ±1.71
DAFNet		**33.01 ±0.20**	**36.54 ±0.21**	**37.23 ±0**	**37.85 ±0.24**	**38.59 ±0**	**39.84 ±0**	**36.35 ±2.38**
Dark [53]	2.5	28.32 ±0.64	-	-	-	31.53 ±0	-	29.39 ±1.60
Graph [5]		17.29 ±1.06	-	-	-	20.25 ±0	-	18.27 ±1.64
Burst [4]		30.80 ±0.28	-	-	-	35.84 ±0	-	32.48 ±2.39
DAFNet		**31.30 ±0.33**	-	-	-	**38.13 ±0**	-	**33.58 ±3.23**
Dark [53]	3	-	-	-	-	30.94 ±0	-	30.94 ±0
Graph [5]		-	-	-	-	12.66 ±0	-	12.66 ±0
Burst [4]		-	-	-	-	34.86 ±0.20	-	34.86 ±0
DAFNet		-	-	-	-	**36.12 ±0**	-	**36.12 ±0**

4.4.2 Influence of Image Quality to Downstream Image Analysis

According to Fig. 4.5, it is hard to differentiate the recovered in-focus images from the ground-truth by human eyes. Another concern is whether the machine also cannot differentiate them, *i.e.*, whether the recovered in-focus images would affect the accuracy of downstream image analysis tasks?

In this subsection, using cell counting that is a typical task of pathology image analysis as an example, we examine the influence of the quality of images yielded by the proposed deep autofocusing approach to the counting accuracy. We utilize a widely used tool *ImageJ*[2] [72] released by National Institutes of Health (NIH) as the

[2]https://imagej.nih.gov/ij/

Table 4.2: The average numbers of counted cells with respect to different ΔD on all samples in Dataset 1.

ΔD	Sample 1		Sample 2		Sample 3		Sample 4		Sample 5		Sample 6		Average	
	Ours	GT	Ours	GT	Ours	GT	Ours	GT	Ours	GT	Ours	GT	Ours	GT
0.5μm	15.17	15.5	42.38	42.13	22.4	22.4	19.4	19.36	18.75	19	22.75	23	22.38	22.43
1μm	12.83	13.16	39.67	40	18.5	19.5	21.12	21.18	18.43	18.86	18.17	17.83	20.25	20.42
1.5μm	14.44	14.56	32	33	33	33	25.5	25.67	21	21	25	25.5	21.78	22.05
2μm	15.33	16	34.5	34	29	28	21	22	23	22	18	17	22.70	22.7
2.5μm	10	10	-	-	-	-	-	-	28	27	-	-	16	15.67
3μm	-	-	-	-	-	-	-	-	27	29	-	-	27	29
Average	14.00	14.27	39.40	39.40	23.44	23.56	20.78	20.84	19.70	19.96	21.20	21.20	21.57	21.69

test platform, which conducts cell counting including the following four steps: 1) gray processing; 2) adjusting brightness and contrast; 3) thresholding; 4) analysis of cell counting.

The in-focus images recovered by our method and the GT in-focus images are taken as input to *ImageJ*, respectively. The results of cell counting are illustrated in Fig. 4.6. It can be found that, the cell counting results on our recovered images are very close to the results on the corresponding GT images. In Table 4.2, we also show the comparison of numbers of counted cells with respect to different ΔD on $Sample1$ to $Sample6$. It can be seen that, compared with the results on the GT, the average cell counting error on our recovered in-focus images is 0.12, which is too small to reduce the accuracy of downstream analysis significantly.

4.4.3 Ablation Study

In this subsection, we provide the empirical ablation analysis about the proposed DAFNet. According to the DAFNet architecture, there are two input defocus images with relative defocus offsets ΔD. Therefore, it is essential to analyze the influence of dual input images and relative defocus offsets to the final performance. Moreover, we

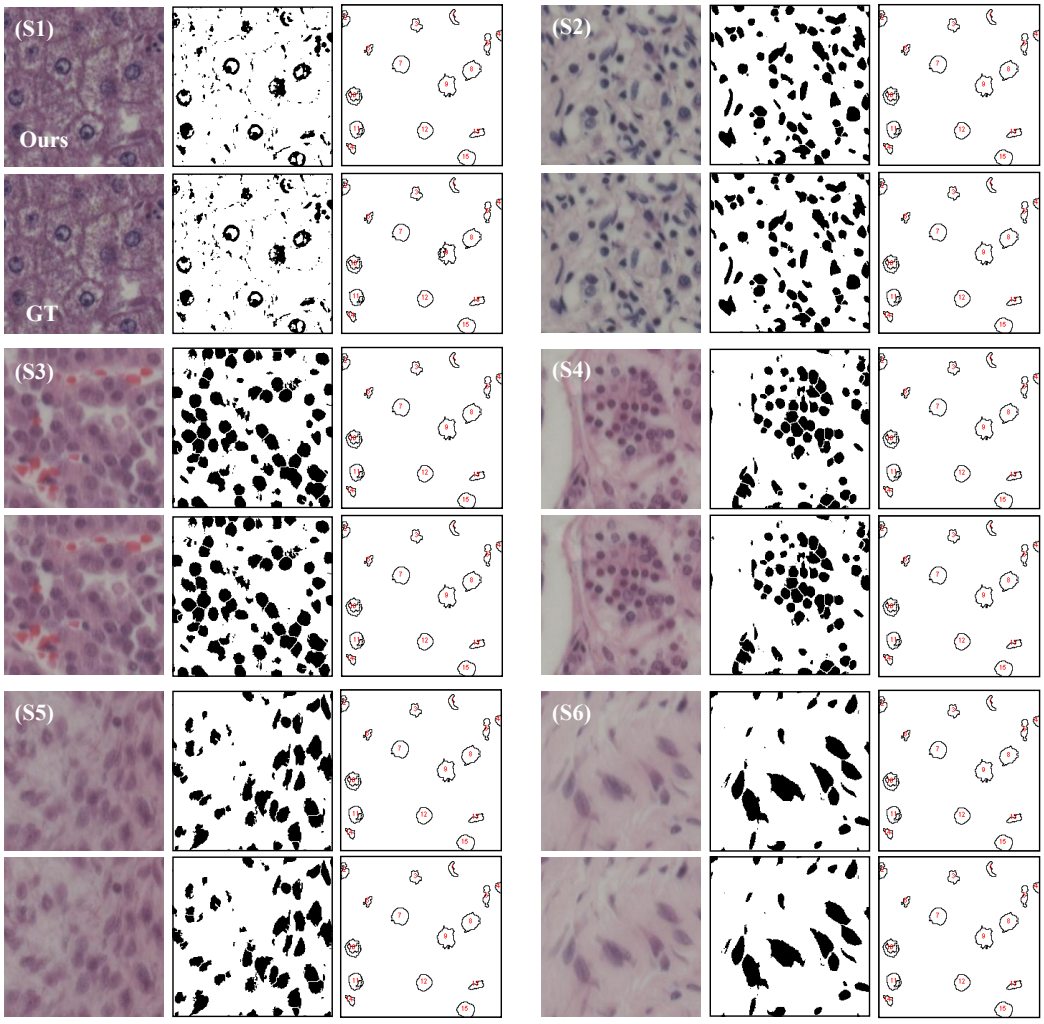

Figure 4.6: Influence of image quality to the accuracy of cell counting. For (S1) to (S6), the cell counting results on our generated image are at the top and the corresponding results of ground-truth are at the bottom. From left to right, the input image for cell counting, the cell segmentation image, and the image of cell outlines counting.

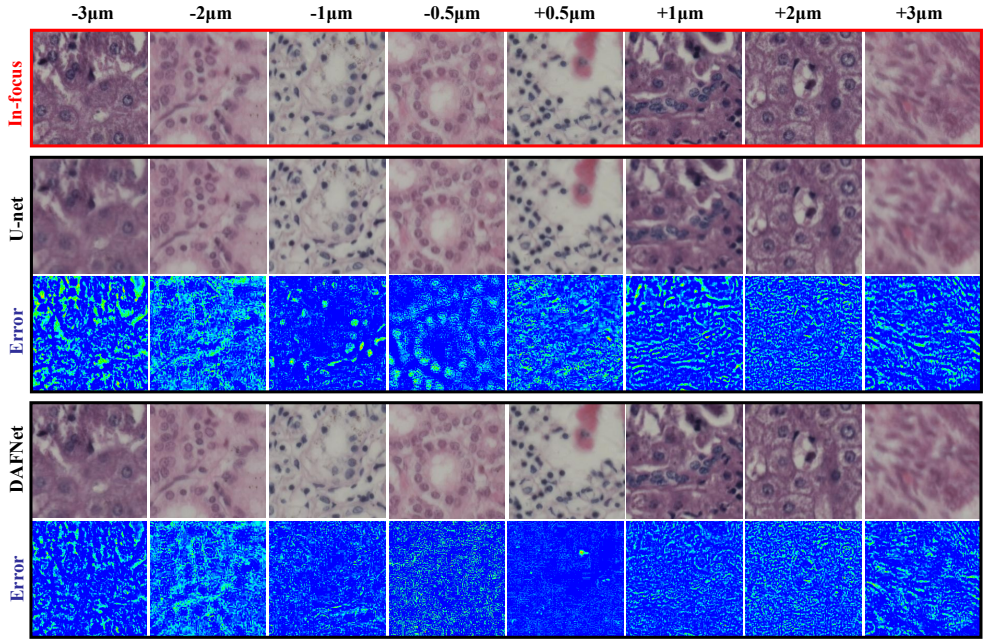

Figure 4.7: Subjective performance comparison on images of Dataset 1. Please enlarge the PDF for more details. The results of U-net and DAFNet, and the corresponding error maps with respect to the groundtruth in-focus images (in red box), are provided.

provide the study of the robustness of the proposed scheme to different test sets. We employ the U-net [69] as the baseline, which takes single input with different relative distance offsets.

Influence of dual input images

In this part, we provide empirical analysis if the dual captured images is really helpful to improve the quality of recovered in-focus images.

Table 4.3 shows objective performance comparison of U-net that takes a single input and our DAFNet that takes Y_1 and Y_2 as inputs. It can be found that, on Dataset 1 and 2, DAFNet achieves much better PSNR performance than U-net for all cases. The average PSNR gains are 2.81dB and 3.49dB over U-net, respectively.

Table 4.3: PSNR performance comparison of U-net and DAFNet on Dataset 1 and Dataset 2 with respect to different ΔD.

Dataset	Methods	Relative Distance Offset ΔD (The mean on the top and standard deviation (SD) on the bottom in each methods)							Average
Dataset 1	U-net	ΔD	-3μm	-2.5μm	-2μm	-1.5μm	-1μm	-0.5μm	39.44 ±3.76
		PSNR	27.96 ±0	33.46 ±4.17	38.28 ±1.28	37.42 ±2.86	38.86 ±1.97	41.61 ±4.08	
		0μm	+0.5μm	+1μm	+1.5μm	+2μm	+2.5μm	+3μm	
		39.44 ±1.58	41.88 ±4.36	38.74 ±2.26	36.05 ±3.11	36.11 ±2.78	32.28 ±3.90	34.88 ±0	
	DAFNet	ΔD	-3μm	-2.5μm	-2μm	-1.5μm	-1μm	-0.5μm	**42.25 ±4.90**
		PSNR	**30.11 ±0**	**33.58 ±2.74**	**38.60 ±1.07**	**38.34 ±1.87**	**39.82 ±1.11**	**47.99 ±1.59**	
		0μm	+0.5μm	+1μm	+1.5μm	+2μm	+2.5μm	+3μm	
		39.61 ±1.32	**48.71 ±1.64**	**39.93 ±1.34**	**37.58 ±2.02**	**36.35 ±2.38**	**33.58 ±3.23**	**36.12 ±0**	
Dataset 2	U-net	ΔD	-3μm	-2.5μm	-2μm	-1.5μm	-1μm	-0.5μm	38.83 ±2.95
		PSNR	33.22 ±0.34	34.82 ±1.74	36.27 ±1.65	37.76 ±1.29	38.61 ±0.99	41.28 ±3.23	
		0μm	+0.5μm	+1μm	+1.5μm	+2μm	+2.5μm	+3μm	
		38.80 ±0.97	40.44 ±3.68	37.71 ±1.54	36.15 ±1.66	35.31 ±1.90	33.49 ±1.81	33.00 ±1.12	
	DAFNet	ΔD	-3μm	-2.5μm	-2μm	-1.5μm	-1μm	-0.5μm	**42.32 ±4.67**
		PSNR	**34.87 ±0.45**	**35.61 ±1.20**	**37.20 ±1.04**	**38.65 ±0.63**	**39.78 ±0.62**	**48.39 ±0.91**	
		0μm	+0.5μm	+1μm	+1.5μm	+2μm	+2.5μm	+3μm	
		39.55 ±0.57	**48.55 ±0.88**	**39.77 ±0.60**	**38.26 ±0.79**	**36.61 ±1.32**	**35.23 ±0.79**	**33.89 ±0.00**	

We also provide subjective performance comparison of U-net and DAFNet in Fig. 4.7 on Dataset 1. For easy assessment, we show the error maps between the recovered in-focus images and the corresponding GT. It can be seen that, compared with U-net, the structure errors produced by DAFNet are smaller, in particular when ΔD is ranging from -1μm to +1μm. Therefore, the proposed DAFNet achieves superior performance than U-net, benefiting from the dual inputs.

Influence of different relative defocus offsets

In this part, we examine the influence of different relative defocus offsets to the final performance.

The PSNR histograms with respect to ΔD on Dataset 1 and Dataset 2 are shown in Fig. 4.9 and Fig. 4.10 respectively. It can be found that: i) For different ΔD, the proposed DAFNet always achieves higher PSNR values than U-net. This demonstrate that the performance of our scheme is robust with respect to ΔD. ii) The highest PSNR gains appear when $\Delta D = +0.5$ μm and $\Delta D = -0.5$ μm. In practical case, most of estimated focal positions also lie in the region of ± 0.5 μm. Therefore, the

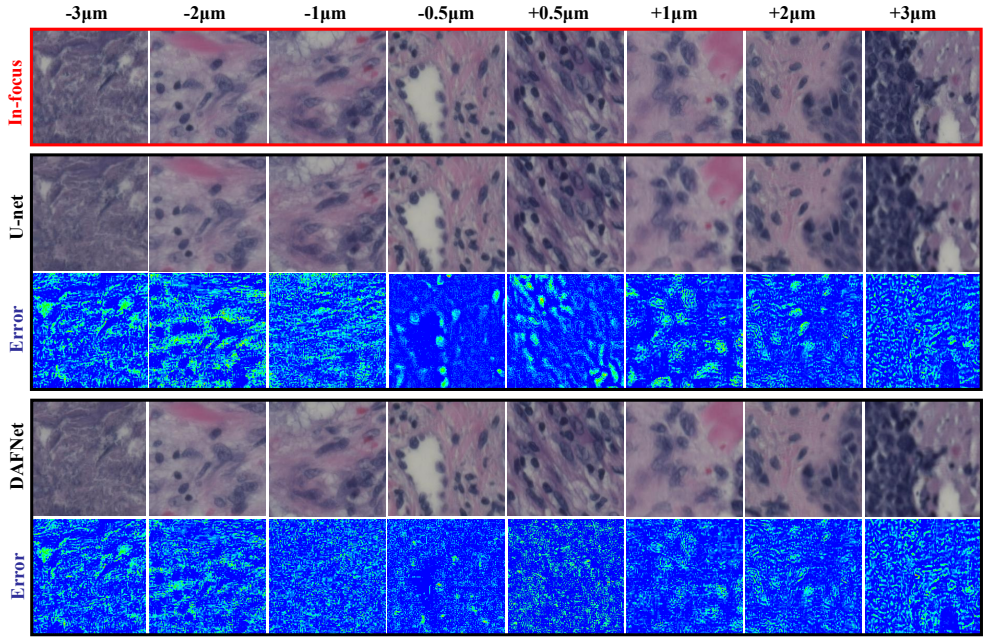

Figure 4.8: Subjective performance comparison on images of Dataset 2. Please enlarge the PDF for more details. The results of U-net and DAFNet, and the corresponding error maps with respect to the groundtruth in-focus images (in red box), are provided.

DAFNet realizes deep autofocusing with high accuracy.

Influence of different test sets

In this part, we examine the robustness of our method to different test sets. In Table 4.3, we provide objective performance evaluation with respect to PSNR on samples of Dataset 2. It can be found that, for test samples from different resources of the training set, our method still achieves the best PSNR performance for all cases. The average PSNR gain over U-net is 3.49dB. The subjective performance comparison of U-net and DAFNet is shown in Fig. 4.8 on Dataset 2. Similar to the results on Dataset 1, the structure errors produced by DAFNet is also much smaller than U-net. These results demonstrate that the proposed DAFNet has a strong generalization

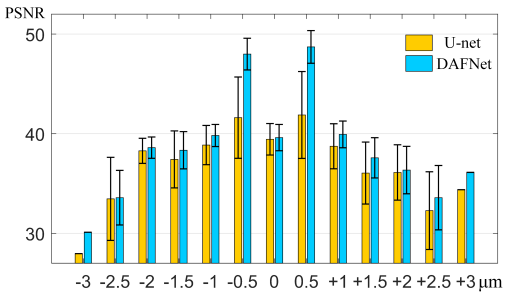

Figure 4.9: PSNR performance comparison of U-net and DAFNet on Dataset 1 with respect to different ΔD.

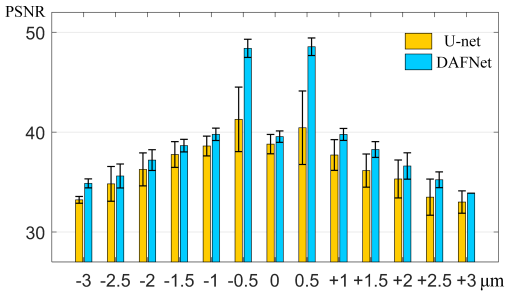

Figure 4.10: PSNR performance comparison of U-net and DAFNet on Dataset 2 with respect to different ΔD.

capability.

4.4.4 Comparison with one-shot deep autofocusing

In this paper, we claim that the proposed scheme performs rapid dual-shot deep autofocusing in WSI. The readers may raise a question: why not design a one-shot deep autofocusing, which would be even faster. Our answer is that, the one-shot manner achieves the highest scanning speed at the cost of imaging quality, while our dual-shot manner achieves a good balance between speed and imaging quality. In this subsection, we conduct experimental analysis to demonstrate this point.

The input of one-shot manner is the captured image with the estimated focal

distance D_0, while our proposed dual-shot strategy takes two images with shifted focal distances $D_0 - \Delta D_1$ and $D_0 + \Delta D_2$ as inputs. For fair comparison, we perform one-shot deep autofocusing using a similar network as DAFNet, but only using the left contracting path and the expansive path of DAFNet.

We assume that the initial plane induced by D_0 is in the center of the thickness of the sample. The PSNR histogram with respect to ΔD on Dataset 1 is shown in Fig. 4.11. The average PSNR of the one-shot scheme is 41.55dB, which is lower than the dual-shot scheme (42.25dB). It can be found that: i) when $\Delta D = 0$, the proposed DAFNet achieves lower PSNR value than the one-shot network. It is worth noting that, both Y_1 and Y_2 are in the range of DoF and the PSNR value of DAFNet is 39.61dB, which means that the image quality is good enough for downstream image analysis. ii) When $\Delta D \neq 0$, the proposed DAFNet always achieves the higher PSNR values than one-shot method. It demonstrates the effectiveness of the dual-shot scheme.

Actually, there is an inevitable error between the initial plane position and the center of the sample thickness. Correspondingly, the probability density distribution of focus points will be shifted along the horizontal axis. Specifically, the initial plane is the baseline of one-shot and dual-shot methods. When the initial plane position changes, the performance (Fig.4.11) with respect to ΔD does not change. Because ΔD is a relative value, which is the distance from the initial plane to the focus point. However, the probability density distribution of focus points with respect to ΔD changes with the error. There is an uneven distribution of focus points of all tiles along the optical axis, and most of the focuses congregate in the center of the sample thickness. For example, when the error is $+0.5\mu m$ (10% for the sample with

Table 4.4: PSNR performance comparison of one-shot and dual-shot methods on Dataset 1 with respect to different Errors (μm).

Methods	Errors	-1 (20%)	-0.5 (10%)	0 -	+0.5 (10%)	+1 (20%)
One-Shot	PSNR	39.15	40.42	41.55	40.62	38.86
	Decline	5.8%	2.7%	-	2.2%	6.5%
Dual-Shot	PSNR	**40.48**	**41.74**	**42.25**	**41.78**	**40.63**
	Decline	**4.2%**	**1.2%**	-	**1.1%**	**3.8%**

a thickness of $5\mu m$), the distribution of focus points moves -0.5μm to the left along the horizontal axis. As stated above, we provide objective performance comparison of one-shot and dual-shot methods with different errors in Tab. 4.4. It can be found that, the dual-shot method is more robust to errors. It is because the dual-shot autofocusing scheme utilizes two tentative possible defocused shooting positions to expand the sensing range of focus points.

In conclusion, the proposed dual-shot scheme achieves the best results in both ideal and practical scenarios. These are the reasons why we adopt a dual-shot manner for network designing.

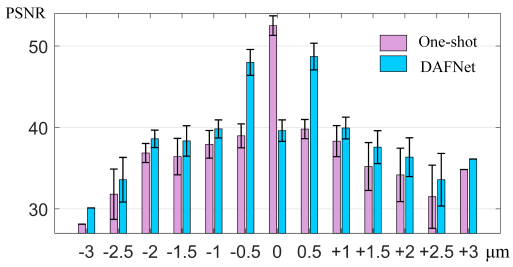

Figure 4.11: PSNR performance comparison of one-shot and dual-shot methods on Dataset 1 with respect to different ΔD.

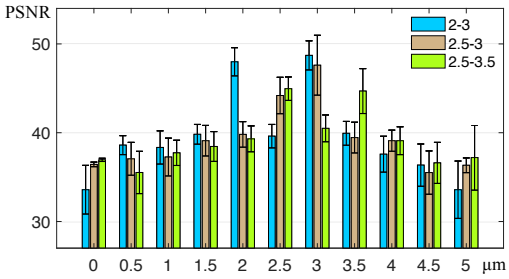

Figure 4.12: PSNR performance comparison of three strategies of DAFNet when Gaussian random variable $n \sim \mathcal{N}(0, 1)$. The label stands for the different absolute shooting positions.

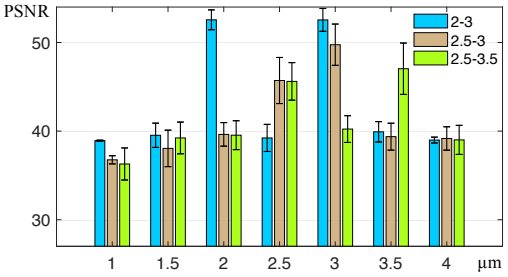

Figure 4.13: PSNR performance comparison of three strategies of DAFNet when Gaussian random variable $n \sim \mathcal{N}(0, 0.5)$. The label stands for the different absolute shooting positions.

4.4.5 Comparison with different dual-shot positions

As stated above, we select two tentative possibly defocused positions near the D_0 as the DAFNet shooting positions. In fact, the focus of the scanned slide is nearly in the center of the tissue (*eg.* the focus position is nearly at $2.5\mu m$ for a sample with a thickness of $5\mu m$). Therefore, the proposed method performs shootings at two absolute positions of $2.5 - \Delta D_1$ and $2.5 + \Delta D_2$. Besides the current version ($\Delta D_1 = \Delta D_2 = 0.5$), we propose two additional shooting strategies: $\Delta D_1 = 0, \Delta D_2 = 0.5$ and $\Delta D_1 = 0, \Delta D_2 = 1$. The corresponding absolute shooting positions are 2 & 3, 2.5 & 3 and 2.5 & 3.5 (μm), respectively. In this subsection, we discuss the DAFNet

performance at different dual-shot positions.

We have converted images in the relative focal distance in the dataset to those in absolute focal distance. We add a Gaussian random variable $n \sim \mathcal{N}(0,1)$ to the dataset and the focus range covers the thickness of the whole sample ($5\mu m$). The PSNR histogram with respect to the sample absolute position is shown in Fig 4.12. The average PSNR of three strategies are 42.25dB, 41.42dB and 40.81dB, respectively. We also add a Gaussian random variable $n \sim \mathcal{N}(0,0.5)$ to the dataset in order to approximate the thinner sample. The PSNR histogram with respect to the sample absolute position is shown in Fig 4.13. The average PSNR of three strategies are 45.72dB, 44.23dB and 42.44dB, respectively. It can be found that: i) In different Gaussian distributions of out-of-focus tiles, the shooting strategy with absolute positions 2 & $3\mu m$ achieves the best objective performance among compared methods. The experiments demonstrate the robustness of the first strategy ($\Delta D_1 = \Delta D_2 = 0.5$) with variational sample thickness. ii) The three strategies achieve the best performance at the shooting positions. For example, the PSNR value of "2-3" is the highest at positions 2 and 3μm evidently. The dual-shot method provides flexibility for the focus distribution of different tiles. In conclusion, these results indicate that the proposed strategy ($\Delta D_1 = \Delta D_2 = 0.5$) makes full use of the dual-shot defocus images, which retain pieces of complementary information about the underlying in-focus image.

4.5 Applications

This section elaborates on the application of the dual-shot virtual autofocusing method in WSI, encompassing the algorithm's workflow, efficiency, as well as its pros and cons.

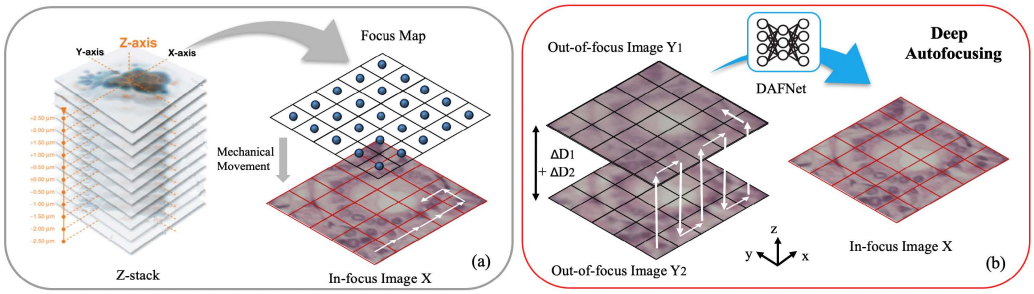

Figure 4.14: Workflow comparisons of the focus map surveying method and the proposed dual-shot deep autofocusing scheme. (a) The conventional focus map surveying method. (b) The proposed dual-shot deep autofocusing scheme. The white arrows mean the scanning order.

The conventional focus mapping technique is depicted in Figure 4.14(a). This traditional approach involves an initial scan to establish a focus map followed by a subsequent scan to capture images in focus. In contrast, our method revolutionizes this process by employing neural networks to eliminate the need for focus mapping, as illustrated in Figure 4.14(b). We introduce a deep learning-driven Whole Slide Imaging (WSI) dual-shot virtual autofocusing network capable of virtually autofocusing defocused tiles from varying distances. This network is adept at generating an in-focus tile directly from two defocused ones, bypassing traditional focus mapping methods. Our method utilizes a computational algorithm to extract the in-focus image, rendering it highly effective for scenarios that demand fast scanning and offline high-precision imaging. The comprehensive WSI workflow is detailed in Appendix A.3.2.

Based on the calculations detailed in Section 3.6, it is evident that executing autofocusing once with the traditional z-stack method incurs a time cost of (10×P + 11×1.4) s. In stark contrast, the deep learning-enhanced WSI dual-shot virtual autofocusing approach introduced in this chapter necessitates merely 2 s for network

inference. This method obviates the necessity for any stepwise motion or gradient calculation, thereby reducing the time required for a single autofocusing instance to (P + 2) s. Comparative analysis reveals that our approach substantially diminishes the time expenditure for autofocusing relative to both traditional z-stack and aberration-guided autofocusing techniques.

Our methodology is characterized by a plethora of benefits, including superior imaging quality, high imaging speed, the elimination of the need for hardware adjustments, cost-efficiency, straightforward adaptability, and the facilitation of offline processing. These attributes render it exceptionally conducive to the high-precision, high-speed, large-scale scanning of pathological slides for offline analysis. However, it's noteworthy that the dual-shot approach might introduce non-directly perceptible FoV errors that could influence network performance. Additionally, the reliance on network inference necessitates the availability of high-performance GPU resources for computational tasks.

4.6 Conclusion

This chapter presents a dual-shot virtual autofocusing method for WSI. Traditional autofocusing techniques rely on constructing focus maps, necessitating repeated mechanical adjustments for refocusing. In contrast, this chapter introduces a deep learning-based WSI virtual autofocusing approach that can generate clear slide images without the need for creating focus maps or utilizing expensive and complex optical path modulations. The proposed method designs a dual-shot virtual autofocusing network, achieving the restoration of in-focus images, reducing the precision requirements of mechanical positioning, and consequently lowering operational costs. Using

high-performance computing devices like GPUs, it enables offline restoration of in-focus images. Experimental results show that this approach achieves superior in-focus restoration quality. Its advantages include high imaging quality, rapid imaging speed, no need for hardware system modulation, cost-effectiveness, ease of transferability, and offline processing capabilities, making it suitable for high-precision scanning and offline processing of pathological slices.

Chapter 5

Semi-blind Deep Restoration of Defocus Images

5.1 Introduction

The method proposed in Chapter 3 predicts the focal position to establish a focus map. Guided by this map, we conduct a second scanning pass for each tile. However, this approach requires simultaneous online system scanning and neural network inference, leading to significant GPU computational costs. The method introduced in Chapter 4 achieves dual-shot in-focus restoration, requiring only a single scanning pass with two shots for each tile. In contrast to the method in Chapter 3, dual-shot in-focus restoration can be efficiently achieved offline using this approach. However, this method still necessitates taking two photos, resulting in relatively high time costs. To expedite the scanning pipeline, can we devise an in-focus restoration method using only a single shot?

The one-shot in-focus restoration method is similar to single-image deblurring

techniques. There are various single-image deblurring methods, including neural networks and postprocessing techniques. For instance, a deep learning-based offline autofocusing method was demonstrated to rapidly and blindly autofocus a one-shot microscopy image [53]. However, this method does not incorporate prior information regarding the imaging system. Furthermore, deconvolution techniques like the Richardson Lucy algorithm necessitate precise prior knowledge of the defocus PSF, which may not always be available [68, 51]. Blind deconvolution methods, which restore images through objective function optimization, can also be employed; however, these methods typically incur high computational costs and are sensitive to factors like image Signal-to-Noise Ratio (SNR) and the selection of hyper-parameters [33, 38].

In this chapter, we introduce a multi-task joint training strategy guided by the PSF prior, where the network simultaneously performs dual predictions for the in-focus image and the defocus image. Our method enables the regeneration of re-defocus images by utilizing estimated blur kernel PSFs. Specifically, due to the effects of aberrations and demosaicing on the PSF, the theoretical Bessel PSF function is unavailable. Therefore, we predict the PSF mask using a neural network rather than traditional optimization algorithms. To address color channel mismatches, we utilize Y-channel data to predict the PSF mask. Subsequently, we generate re-defocus images convoluted by the corresponding PSF from classification. Finally, the network imposes joint constraints on both in-focus and defocus images, thereby significantly enhancing image restoration performance. Experimental evaluations underscore the advantages of this method, including high scanning efficiency, elimination of the need for hardware adjustments, cost-effectiveness, easy portability, and the option for offline processing.

5.2 Preliminaries and Motivations

Objectively, achieving in-focus restoration directly from a defocus image is more challenging than blind deblurring of general images. Defocus leads to unavoidable aberrations, and subsequent demosaicing further reduces image quality.

5.2.1 Defocus Aberrations

In microscope imaging, the effects of defocus typically manifest as a decline in image quality, specifically in terms of reduced resolution, weakened contrast, and the possible emergence of specific aberrations. The detailed aspects of defocus effects include:

Resolution Decline

During defocus, the PSF widens, leading to the loss of image details. This means that two points that could be distinguished close together become blurred and indistinguishable, lowering the effective resolution of the microscope.

Contrast Weakening

Defocus also results in the decline of image contrast, making the image appear flatter, lacking depth and detail. In biological microscope imaging, this could make the observation of cellular structures more difficult.

Emergence of Specific Aberrations

Defocus may exacerbate or introduce specific aberrations, such as spherical aberration, coma, and chromatic aberration:

- **Spherical Aberration**: This is a result of the lens shape (spherical) causing light rays from different positions to focus at different points, resulting in different levels of sharpness between the center and the edges of the image. Under defocus conditions, spherical aberration becomes more pronounced because the difference in the focal points of the light rays increases;

- **Coma**: Coma is a type of optical aberration that occurs when light enters the lens at an angle off the optical axis. In microscope imaging, especially when out of focus, coma can cause an asymmetric spread of the PSF on the imaging plane, manifesting as elongation in a specific direction (usually towards or away from the optical axis);

- **Chromatic Aberration**: Chromatic aberration is usually related to the dispersion of wavelengths, but in the case of defocus, light of different colors, due to different focal points, may also produce different diffusion patterns on the imaging plane, further affecting image quality.

Therefore, achieving in-focus restoration from defocus images presents significant challenges due to the inherent complexity of the aberrations introduced by defocus. The trio of resolution decline, contrast weakening, and the emergence of specific aberrations such as spherical aberration, coma, and chromatic aberration due to defocus pose substantial obstacles in in-focus restoration.

5.2.2 Demosaicing

Definition and Purpose

Demosaicing is an image processing technique used to reconstruct a full color image from the raw image data captured by a digital camera sensor with a Color Filter Array (CFA). The most common CFA is the Bayer filter, which arranges red, green, and blue filters in a certain pattern, with each pixel capturing information from only one color channel. Demosaicing algorithms reconstruct information for all three color channels for each pixel by interpolating the color values of neighboring pixels. Executing demosaicing correctly is challenging because it requires accurately estimating the missing color information while minimizing the artifacts and color distortions caused by interpolation.

Image Degradation Caused by Defocus and Demosaicing

- **Defocus Aberrations Reduce Image Quality**: Demosaicing relies on high-quality raw image data to reconstruct color information. If the raw image is blurred due to defocus, the lack of sharpness and detail in the image data can lead to inaccurate interpolation results, further reducing the quality of the final image;

- **Mismatch of Color Channels Leads to Artifacts and Distortions**: Defocus makes it more difficult to accurately estimate the missing color information during the demosaicing process, thereby increasing color distortions and artifacts. Defocus causes the light rays in the image to focus inaccurately on the sensor, and aberrations caused by different wavelengths further affect the focus

state of different color channels. Specifically, due to the dispersion effects of light, different colors (wavelengths) have different refractive indices in the lens system, which means that red, green, and blue light focus at different positions after passing through the same optical system. On the image sensor, this can lead to different levels of defocus for each color channel, thereby affecting image quality and color accuracy.

Therefore, achieving accurate color reconstruction is particularly challenging when the raw image is blurred due to defocus. Defocus not only diminishes image sharpness and detail, which are vital for precise interpolation but also introduces aberrations that cause different color channels to focus at varying positions on the sensor. As a result, defocus complicates the demosaicing process, potentially leading to color distortions and artifacts that degrade the overall image quality.

5.3 The Proposed Method

To address the challenge of in-focus image restoration from defocus images with unavoidable aberrations and mismatched demosaicing, this study proposes a neural network designed to recover in-focus images from those exhibiting different degrees of defocus for semi-blind deep restoration.

5.3.1 Deep Restoration of One-shot Autofocusing Method

This chapter proposes a multi-task approach, where the network simultaneously performs dual predictions for both in-focus images and defocus features. In the semi-blind reconstruction determined in the PSF section, we select the defocus distance

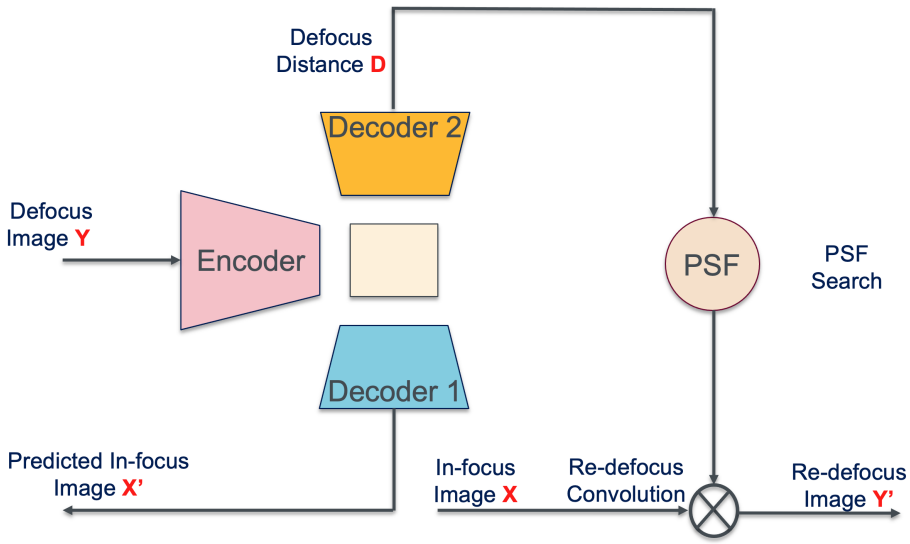

Figure 5.1: Semi-blind deep restoration of one-shot autofocusing method

as the defocus feature to guide the acquisition of the PSF. The methods contain the following parts:

- **Encoder**: the input is defocus image Y and the encoder extract the defocus features;
- **Decoder1**: the decoder1 predicts in-focus image X';
- **Decoder2**: the decoder2 predicts defocus distance D;
- **PSF Search**: the PSF can be searched by the corresponding defocus distance. We achieve PSF at different defocus distances by in-focus & defocus pairs network prediction;
- **Re-defocus**: the final output Y' is re-defocus image convoluted by in-focus image X and PSF.

5.3.2 The Semi-blind Deep Restoration Network

To predict aberration-free in-focus images, we develop an encoder-decoder structure complemented by a multi-task learning approach leveraging the U-Net architecture, as shown in Fig.5.2. The first half of the network is dedicated to feature extraction, while the latter half focuses on image restoration and defocus distance prediction. This architecture is tailored for the extraction of multiple features (defocus distances and images) across a range of defocus conditions. The network implements a max-pooling operation to condense the feature maps, effectively halving their spatial dimensions. As the data proceeds to the backend segment—dedicated to either in-focus restoration or distance prediction—it undergoes a transposed convolution operation, which expands the spatial dimensions of the feature map back to their original size. Ultimately, the image branch of the network generates a predicted in-focus image, whereas the distance branch delivers a defocus distance prediction. This prediction is represented as an integer value, determined through a distance classification process that utilizes a softmax operation with six categories from 0 to $5\mu m$. The detailed formula for the deep network model is as follows:

$$X = F\left(Y_{\Delta D}, \theta\right) \tag{5.3.1}$$

where ΔD stands for defocus distance from 0 to $+5\mu m$ with $1\mu m$ step, F is in-focus restoration network, Y stands for the input (defocus images), X means output (predicted in-focus images), and θ means network parameters.

In the PSF dictionary, we designed PSFs for six different defocus distances, ranging from 0 to $5\mu m$ with $1\mu m$ intervals. Based on the results of distance classification,

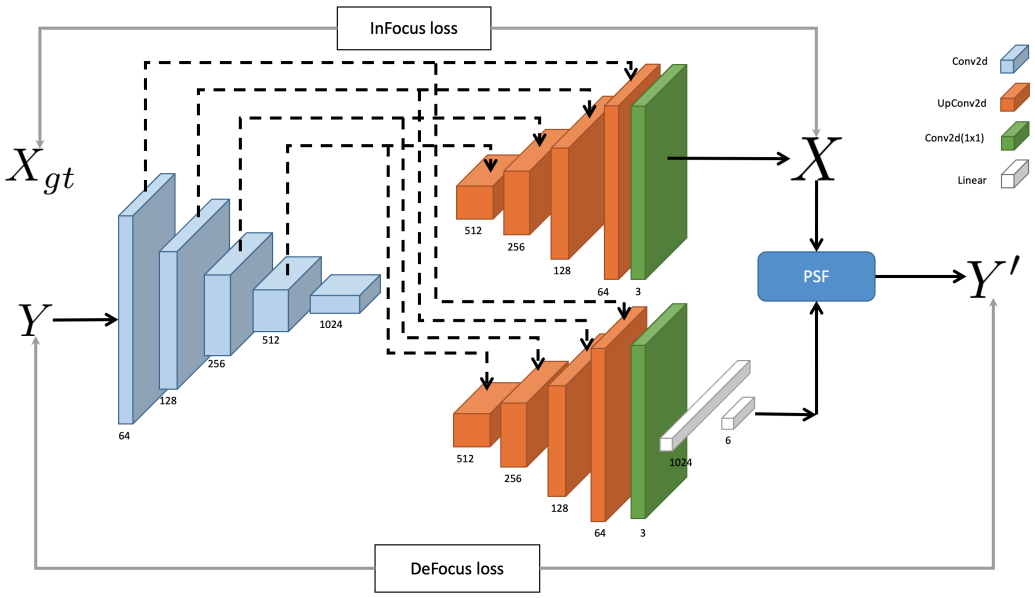

Figure 5.2: Semi-blind deep restoration network for one-shot autofocusing (OAF-Net) of digital pathology.

the corresponding PSF is selected. Training of the network revealed that discrete PSFs do not allow for gradient feedback. Therefore, we devised a method involving mixed PSFs: by taking the probabilities of the six categories from the distance classification and performing a weighted sum with the corresponding PSFs. Under this mixed PSF design, network gradients are effectively fed back.

5.3.3 The PSF Restoration Network

We can predict the blur kernel function, i.e., the PSF, simply by using a pair of in-focus and defocus images, as shown in Fig.5.3. The input to the neural network is the in-focus image, from which the blur features are extracted through the designed convolution, and the predicted PSF is obtained after the reshape operation. By convolving the input clear image with the predicted PSF, a re-defocus image can be

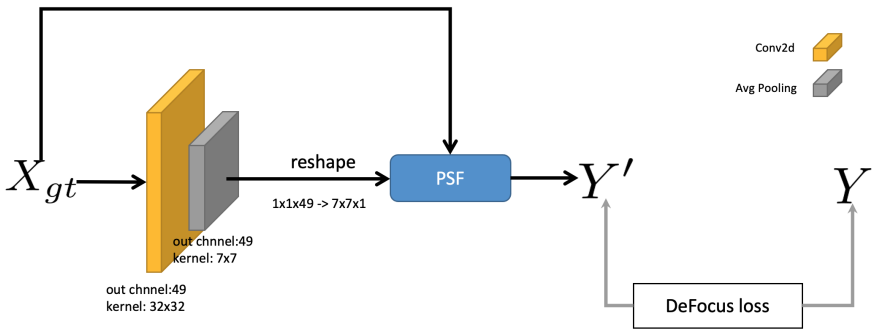

Figure 5.3: PSF restoration network.

obtained. The loss calculation between the re-defocus image and the GT defocus image completes the design of the entire network.

To avoid mismatched demosaicing, we use the YUV image channels instead of the traditional RGB channels. Under the influence of defocus aberrations, the demosaicing of different color channels cannot match. By designing the network using the intensity Y channel, we can effectively avoid the strict requirements on color in the RGB channels and obtain a realistic and objective PSF.

5.3.4 Dataset and Network Training

Dataset

The method employs the open-source synthetic pathological database [17], for training the in-focus restoration network. We utilize the GT image to generate defocus image and create the z-stack. Each z-stack consists of 6 images with varied defocus distances, with a defocus step of 1 μm and a range from 0 to +5 μm. We design a simulated PSF model similar to WSI system with lens 0.75 NA/ 20×. The dataset we generated contains 5000 z-stacks. Our ratio of training set, validation set and test set is 7:2:1.

In the real dataset, we still utilize the dataset collected by Jiang *et al.* [30].

Semi-blind Deep Restoration Network Training

The semi-blind in-focus restoration network employs the ADAM optimizer and sets the learning rate at 0.001, specifically aiming at gradient descent optimization for the MSE loss. The network underwent 10 training epochs on an NVIDIA GTX3090Ti graphics card, with a batch size of 32 for each epoch. In total, the training process spanned approximately 9 hours, and the overall size of the network model stands at 1.4G.

Loss Function

In this part, we design two experiments for comparative studies.

- **Baseline**: We design a baseline network called **In-focus Loss** method in Exp1. The network is a regular U-net, containing an encoder and a decoder only. The network output predicted the in-focus image directly. In this scenario, we utilize MSE loss as the loss function:

$$L_I = L_I^{MSE} \tag{5.3.2}$$

- **Ours**: We design an image-image joint network with classification search called **In-focus & Defocus Loss** method. The network contains an encoder and two decoders separately. The network output predicted the in-focus image and the corresponding defocus distance. We could implement PSF search by the classification of predicted defocus distance. We design a joint loss, containing

the in-focus image loss and defocus image loss (classification). In this scenario, we utilize MSE loss as the loss function:

$$L_{ID} = \alpha \cdot L_I^{MSE} + (1 - \alpha) \cdot L_D^{MSE} \tag{5.3.3}$$

5.4 Experiments

5.4.1 Performance Comparison on Dataset1

In this section, methods (Exp.1-2) are employed for performance comparison on Dataset1 [30]:

- **In-focus Loss**: regular U-net containing an encoder and a decoder only, abbreviated as **Baseline**;
- **In-focus Loss & Defocus Loss**: an image-image joint network with PSF searched by classification of defocus distance, abbreviated as **Ours**.

In-focus restoration results at different defocus distances as illustrated in Fig.5.4. Our method only introduces an additional decoder and designs multi-task joint constraints. Our method does not rely on the network itself, so there is no need to compare it with other SOTA neural network methods.

Experimental results elucidate the following insights: **Ours** outperforms **Baseline** markedly in terms of focus restoration. The introduced multi-task strategy, incorporating in-focus / defocus loss constraints, adeptly manages focus restoration across varying defocus distances. Subjectively speaking, the outcomes from **Ours** satisfactorily fulfill the autofocusing requirements for pathological data.

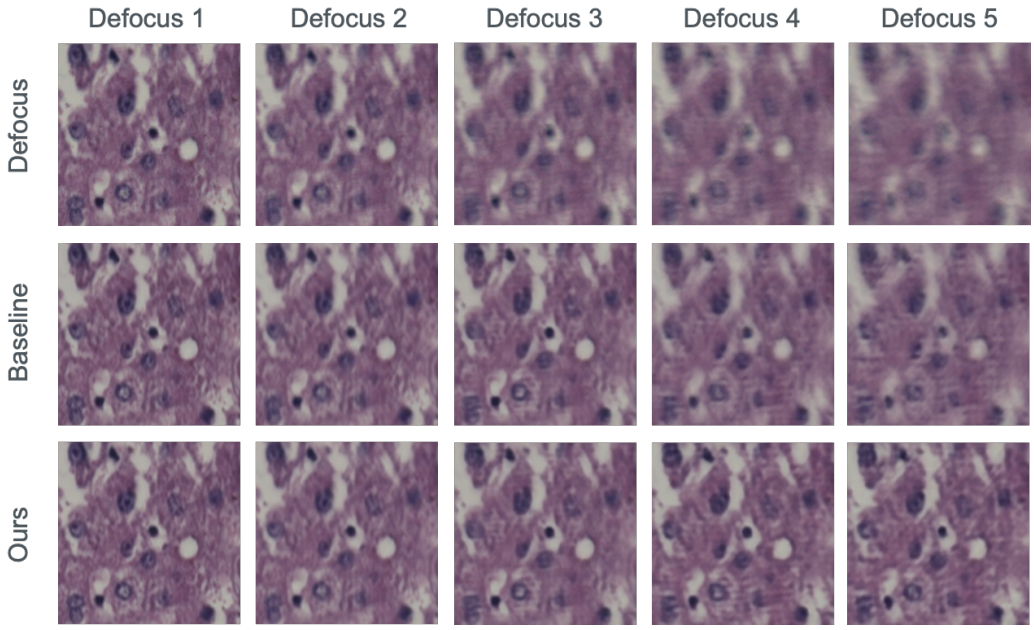

Figure 5.4: Subjective performance comparison of in-focus restoration methods

To objectively assess the efficacy of in-focus image restoration, this study utilizes PSNR and SSIM as the primary metrics for evaluation. In the PSNR metric, **Ours** registers a score of 36.04 dB, marking a 2.7% enhancement compared to **Baseline**, which scores 35.09 dB. In terms of SSIM, **Ours** achieves 0.9497, improving by 0.9% over **Baseline**'s score of 0.9411. Objective comparisons from experimental data affirm the superior performance of our method.

Differential performance across various defocus distances is illustrated in Fig. 5.5. (1) The experimental findings indicate that **Ours** secures the best performance at diverse defocus distances. (2) As shown in Fig. 5.5, even though the efficiency of in-focus restoration diminishes with an increase in defocus distance, **Ours** still maintains a performance of 33.34 dB at the maximum defocus distance of 5μm. Consequently, our approach demonstrates unparalleled effectiveness across the entire spectrum of defocus distances.

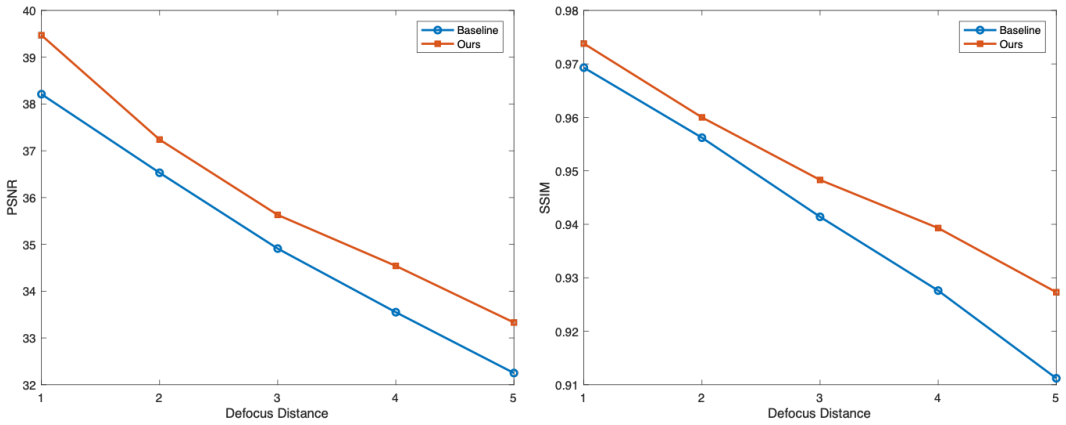

Figure 5.5: Objective comparison of in-focus recovery performance at different defocus distances. Left: PSNR, Right: SSIM

5.4.2 Performance Comparison on Dataset2

To assess the generalization of the proposed deep restoration method, a real-data analysis was performed on Dataset2 [30].

In-focus restoration results at different defocus distances on Dataset2 as illustrated in Fig.5.6. Experimental results elucidate the following insights: **Ours** outperforms **Baseline** still markedly in terms of focus restoration. The introduced multi-task strategy, incorporating in-focus / defocus loss constraints, adeptly manages focus restoration across varying defocus distances. Subjectively speaking, the generalization of **Ours** satisfactorily fulfill the autofocusing requirements for pathological data.

To objectively assess the efficacy of in-focus image restoration, this study utilizes PSNR and SSIM as the primary metrics for evaluation. In the PSNR metric, **Ours** registers a score of 35.11 dB, marking a 5.6% enhancement compared to **Baseline**, which scores 33.26 dB. In terms of SSIM, **Ours** achieves 0.9523, improving by 1.7% over **Baseline**'s score of 0.9375. Objective comparisons from experimental data affirm

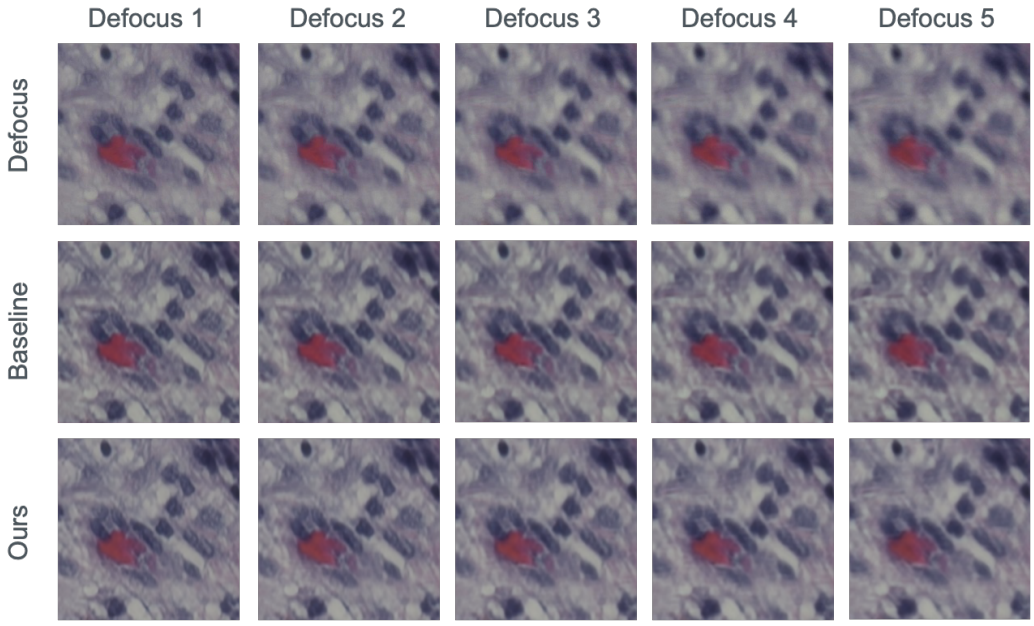

Figure 5.6: Subjective performance comparison of in-focus restoration methods

the superior performance of our method.

Differential performance across various defocus distances is illustrated in Fig. 5.7. (1) The experimental findings indicate that **Ours** secures the best performance at diverse defocus distances. (2) As shown in Fig. 5.7, even though the efficiency of in-focus restoration diminishes with an increase in defocus distance, **Ours** still maintains a performance of 32.34 dB at the maximum defocus distance of 5μm. Consequently, our approach demonstrates unparalleled effectiveness across the entire spectrum of defocus distances.

5.5 Applications

This section elucidates the application of the one-shot virtual autofocusing method for WSI, encompassing the algorithm workflow, its operational efficiency, and its merits

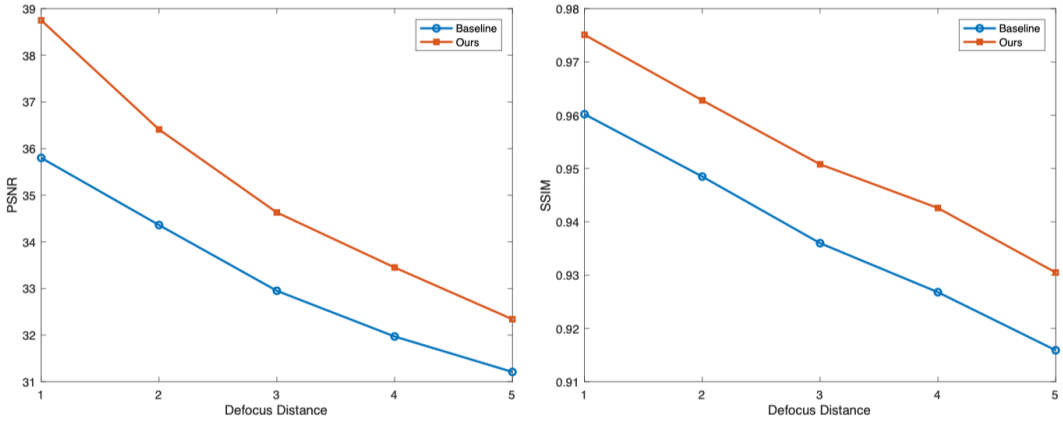

Figure 5.7: Objective comparison of in-focus recovery performance at different defocus distances. Left: PSNR, Right: SSIM

and drawbacks.

The conventional focus map plotting method is illustrated in Fig.5.8(a) as per reference [7]. However, in our approach, processes involving the focus map (including the initial scan to construct the focus map and the second scan to capture in-focus images) are superseded by the neural network, as portrayed in Fig.5.8(b). Our method introduces a deep learning-based WSI one-shot virtual autofocusing network to virtually autofocus defocus tile images at various defocus distances, directly converting a single defocus tile into an in-focus one. Distinct from the focus map measurement technique, our approach retrieves in-focus images via computational algorithms, making it apt for high-efficiency scanning scenarios. The detailed WSI workflow is shown in Appendix A.3.3.

From the calculations presented in Section 3.6, it is evident that utilizing the traditional z-stack method for a single autofocusing execution incurs a time cost of $(10\times P + 11\times 1.4)$ s. However, this chapter introduces a deep learning-based WSI one-shot virtual autofocusing method that, through network inference, reduces the

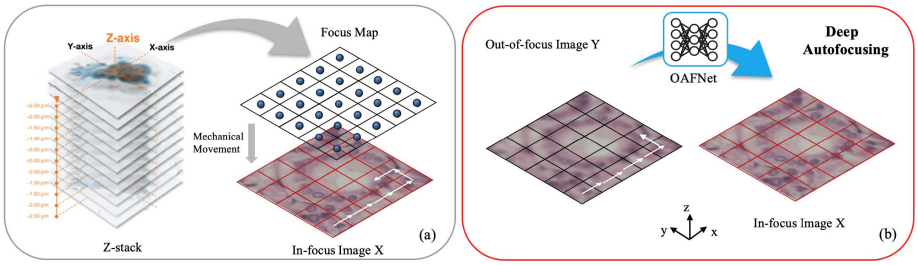

Figure 5.8: Autofocusing comparisons of the focus map surveying and the dual-shot deep autofocusing. (a) The conventional focus map surveying method, (b) The one-shot deep autofocus OAFNet

measurement time to just 4.8 s. This approach eliminates the need for any step-wise mechanical movements and gradient computations, requiring only a single scan. Thus, the time cost for executing autofocusing once using our method is significantly reduced to 4.8 s. The calculations demonstrate that compared to the traditional z-stack method and aberration-guided autofocusing techniques, our method significantly decreases the autofocusing time cost. Furthermore, when compared to the dual-shot autofocusing method, our approach offers a doubling in scanning efficiency.

Our method boasts exceptionally high scanning efficiency, negates the need for hardware system modulation, offers low costs and ease of transport, and provides offline processing capabilities. It is exceptionally suited for the high-speed, large-scale scanning and offline processing of pathological slides. On the limitations front, our approach requires a semi-deterministic knowledge of the PSF. Moreover, the inference process of the network necessitates computational resources supported by high-performance GPUs.

5.6 Conclusion

This chapter introduces a one-shot WSI virtual autofocusing method based on deep learning. The algorithm takes defocus blurred images as inputs and directly produces the corresponding in-focus images, theoretically replacing the traditional approach of mechanically scanning and adjusting focus for each tile. Given the varied degrees of image blurriness at different defocus distances, an OAF-Net with defocus attention is proposed. We propose a multi-task joint training approach, where the network simultaneously performs dual predictions for both in-focus images and defocus distances. Following this, an in-focus restoration and distance prediction multi-decoder is developed. Finally, both subjective and objective experimental results affirm that the in-focus restoration network designed in this study can swiftly process defocus blurred images and achieve in-focus restoration outcomes without compromising image quality. Boasting high-throughput, high-speed imaging, no need for hardware system modulation, cost-effectiveness, ease of migration, and offline processing capabilities, this method is especially suited for high-speed, large-volume scanning and offline processing of pathological slides.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This article is focused on the research of autofocusing methods for whole slide digital pathology imaging and aims to achieve high-speed, high-precision imaging of the entire FoV of pathological tissue slices. The primary research work and conclusions are as follows:

- **Aberration-aware Focal Distance Prediction**: To address the slow focusing speed caused by existing z-stack focus estimation methods, a WSI autofocusing method guided by aberrations is proposed. Due to the different refractive indices of the sample medium, spherical aberration occurs at the defocus position of the medium interface, ultimately leading to asymmetric features in positive and negative defocus images. This method judiciously leverages this negative characteristic of asymmetric imaging to construct a defocus classification network. This classifies samples with different feature properties into

positive / negative categories. Through the classification network, samples with similar features are obtained, which further boosts the performance of subsequent refocusing networks. This method effectively learns the mapping between the defocus image and the defocus distance. Overcoming the limitations of traditional z-stack focus estimation, accurate distance prediction is achieved with a single estimation, and it's compatible with current WSI methods. It is suitable for focus prediction, such as focus map generation.

- **Dual-shot Deep Autofocusing with a Fixed Offset Prior**: In response to the issue of uneven focus distribution in pathological samples leading to poor imaging quality, a virtual autofocusing method for WSI with two shots based on deep learning is presented. The dual-shot method provides a fixed offset prior, reducing the fitting difficulty of the inverse process of the imaging model. Considering the large focus distribution range, unevenness, and inconsistency of pathological samples, the neural network design for the dual-shot method integrates complementary information from input images at two different focal lengths with fixed positions, breaking the constraints of uneven distribution and DoF in pathological samples. Compared to the one-shot method, the dual-shot method retains advantages such as hardware virtualization, cost-effectiveness, and offline processing capabilities. Without sacrificing scanning efficiency, it achieves superior imaging quality and rapid, high-precision virtual refocusing. This method is suitable for high-precision imaging scenarios, obtaining restored in-focus images through computer offline processing.

- **Semi-blind Deep Restoration of Defocus Images**: Addressing the inefficiency of existing focus map scanning strategies, a deep restoration method

for WSI with one-shot is proposed. Compared to direct end-to-end approaches, this method proposes a multi-task joint training approach, where the network simultaneously performs dual predictions for both in-focus images and defocus features. We can achieve re-defocus images by the semi-blind deep restoration with a classification PSF search with a PSF mask generation. The network can impose joint constraints on both in-focus and defocus images, thereby significantly enhancing image restoration performance. This method only requires a single shot at any defocus distance, fundamentally avoiding repeated focusing movements and camera exposure processes. Successfully virtualizing hardware functions, this approach boasts high throughput, speed, cost-effectiveness, practicality, and offline processing capabilities, significantly enhancing scanning efficiency. It's suitable for high-efficiency scanning scenarios, with computer offline processing to obtain restored in-focus images.

6.2 Future Work

Although advancements have been made in the autofocusing methods for WSI in this study, there are still some limitations:

- **Sample Range Limitation**: The pathological slide samples used in this study might not cover all types of pathological slides, such as some rare ones with unique properties.
- **Hardware Dependence**: While the proposed method has advantages at the algorithmic level, it might be constrained by specific imaging hardware, necessitating appropriate algorithm adjustments.

- **Processing Speed**: Even though the current method is relatively efficient, further speed enhancements are required for some high-demand real-time applications, such as real-time remote medical diagnosis.
- **Generalization Issues**: The deep learning method proposed in this study performs well on specific datasets, but its generalization to other imaging devices still needs further validation.
- **Computational Resource Requirement**: Utilizing deep learning may demand significant computational resources, especially on vast amounts of pathological slide data, which might limit its application in low-end devices or constrained environments.

Given these limitations, future research will need to further optimize and improve, ensuring the broad applicability and efficiency of autofocusing for WSI. Based on this research and its conclusions, the following prospects for future work are outlined:

- **Optimization and Deepening**: Although the aberration-guided autofocus method has demonstrated its speed and accuracy, further optimization and verification are essential under broader pathological slide types and imaging conditions. Specifically, some special pathological slides, such as those with complex backgrounds or low contrast, might need further optimization.
- **Hardware Integration**: The current method has achieved commendable results at the algorithmic level, but integrating it with modern digital pathology imaging hardware is a worthwhile research direction. A deep fusion between hardware and algorithms might further boost scanning and imaging speeds while reducing costs.

- **Adaptive Methods**: Given the potential non-uniform distribution of focus in pathological slides, future studies can explore more adaptive focusing strategies, dynamically adjusting focusing strategies and parameters based on the specific characteristics of the slide for improved imaging quality.

- **Expanding Application Range**: While this study primarily focuses on WSI, the core concepts might be applicable to other medical imaging techniques, such as MRI, CT, or ultrasound. Future work could consider expanding into these domains.

- **Multi-modality Fusion**: Considering that pathology might involve various imaging modes, like bright field and fluorescence, future studies can look into effectively merging these different imaging modes, leveraging their individual strengths for enhanced imaging quality and diagnostic accuracy.

- **Online Real-time Processing**: Although the current method prioritizes offline processing, with the advancement of computational capabilities, there's potential to develop online, real-time imaging and focusing strategies to meet the needs of real-time diagnosis and remote medicine.

In summary, this study offers new and effective autofocusing strategies for WSI, yet there remain many avenues worth exploring in the future to achieve higher imaging quality and scanning efficiency.

Appendix A

Appendix

A.1 Defocus Degradation Imaging Model

We derive the defocus degradation imaging model in the following in general imaging systems. Firstly, we assume an ideal point light located in the optical axis with distance D away from the thin lens. The complex amplitude distribution of point light can be expressed by

$$U_0(x,y) = \exp\left(ik\sqrt{x^2+y^2+D^2}\right), \tag{A.1.1}$$

where $k = 2\pi/\lambda$ stands for the wavenumber.

Next, the light propagates through the lens with a phase delay, *i.e.*, the transmission function, which is defined as

$$t(x,y) = P(x,y)e^{i\phi(x,y)} = P(x,y)\exp\left[-i\frac{k}{2f}\left(x^2+y^2\right)\right], \tag{A.1.2}$$

where $P(x, y)$ is pupil function as

$$P(x,y) = \begin{cases} 1, \sqrt{x^2+y^2} \leq R \\ 0, other \end{cases} . \tag{A.1.3}$$

Considering the additional optical aberrations, the corresponding phase delay is given by

$$t'(x,y) = \exp\left[jk\left(n(\lambda)-1\right)\Delta(x,y)\right], \tag{A.1.4}$$

where $n(\lambda)$ is refraction index changing with wavelength, and Δ is Zernike basis which contains orthogonal polynomials on the unit disk [59]. The electric field after lens is the product of point light electric field and transmission function

$$U_1(x,y) = t(x,y)t'(x,y)U_0(x,y). \tag{A.1.5}$$

Then the field propagates from the lens to the sensor with the transfer function [21]

$$H_{\mathrm{s}}\left(f_x, f_y\right) = \exp\left[iks\sqrt{1-\left(\lambda f_x\right)^2-\left(\lambda f_y\right)^2}\right], \tag{A.1.6}$$

where (f_x, f_y) are spatial frequencies. In the Fourier domain, the linear relationship in transfer model is given by

$$\mathcal{F}\left\{U\left(x', y'\right)\right\} = \mathcal{F}\left\{U_{\,1}\left(x, y\right)\right\} \cdot H_{\mathrm{s}}\left(f_x, f_y\right), \tag{A.1.7}$$

where $\mathcal{F}$ stands for 2D Fourier transform. We know the camera measures optical intensity, *i.e.*, the magnitude-squared of complex amplitude distribution. Finally, the

final PSF is given by

$$\mathrm{PSF}_D(x, y; D) = \left|\mathcal{F}^{-1}\left\{\mathcal{F}\left\{t \cdot t' \cdot U_0\right\} \cdot H_\mathrm{s}\right\}\right|^2 (x, y). \tag{A.1.8}$$

According to the expression of PSF, it is hard to settle this inverse problem from formula. It is necessary to note that the autofocusing task in WSI is to estimate defocus distance, and shift platform to the position. Therefore, it is feasible to estimate the defocus distance with a neural network.

A.2 Uniformity of Dual Focus

In the one-shot defocus image restoration methods, as illustrated in Figure A.1, the simplified single-layer imaging model can be expressed as

$$f * h\left(\Delta_{D_1}\right) = g_1(x, y), \tag{A.2.1}$$

where f means in-focus images GT, h stands for PSF, $\Delta_{D_1} = D_1 - D$ is defocus distance and g_1 means defocus image 1.

From the equation, it can be seen that g_1 is a known variable, while f and h are unknown variables. Hence, the one-shot defocus image restoration method corresponds to a blind deconvolution problem. Similarly, introducing the second imaging model

$$f * h\left(\Delta_{D_2}\right) = g_2(x, y), \tag{A.2.2}$$

where $\Delta_{D_2} = D_2 - D$ is defocus distance 2 and g_2 means defocus image 2.

In actuality, the second shot is obtained by moving the objective lens from the

position of the first shot, that is, the defocus distance Δ_{D_2} is predetermined and known, which can be denoted as $\Delta_{D_2} = \Delta_{D_1} + \Delta$, where Δ is defined by us. Converting the above two image models to the frequency domain can be represented as

$$F \cdot H\left(\Delta_{D_1}\right) = G_1, F \cdot H\left(\Delta_{D_1} + \Delta\right) = G_2, \tag{A.2.3}$$

where F denotes the spectrum of the in-focus image, H represents the Optical Transfer Function (OTF), G_1 indicates the spectrum of the defocused image g_1, and G_2 represents the spectrum of the defocused image g_2. Combining the above expressions, we have:

$$\frac{H\left(\Delta_{D_1} + \Delta\right)}{H\left(\Delta_{D_1}\right)} = \frac{G_2}{G_1}, \tag{A.2.4}$$

In academic literature, the mathematical expression of the Airy model for the defocused PSF in the frequency domain, denoted as OTF, is described as follows [8]

$$OTF(u, v, \mathrm{NA}, \lambda, \Delta_D) = \frac{2J_1\left(2\pi \mathrm{NA} \frac{\sqrt{u^2+v^2}}{\lambda} \Delta_D\right)}{2\pi \mathrm{NA} \frac{\sqrt{u^2+v^2}}{\lambda} \Delta_D}, \tag{A.2.5}$$

where u, v represents the spatial frequency coordinate in the frequency domain, NA denotes the numerical aperture, λ stands for the wavelength, Δ_D signifies the defocus distance of the objective lens, and J_1 is the Bessel function of the first kind.

By substituting the OTF into the equation and simplifying the constant terms, we obtain

$$J_1\left(\Delta_{D_1} + \Delta\right) = K\left(1 + \frac{\Delta}{\Delta_{D_1}}\right) J_1\left(\Delta_{D_1}\right). \tag{A.2.6}$$

Performing divisions involving Bessel functions can lead to intricate expressions,

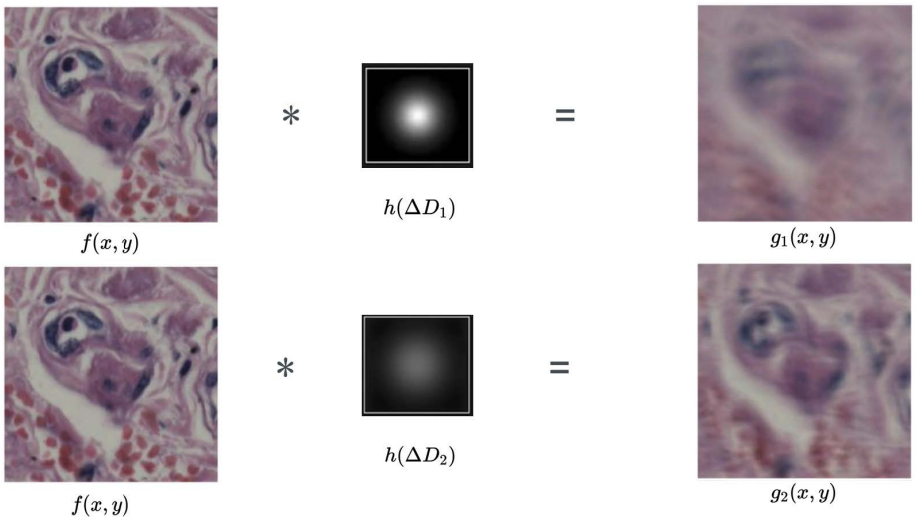

Figure A.1: Schematic diagram of two defocus shots

especially when considering Bessel functions of a real order. Generally, calculating divisions with Bessel functions may necessitate the use of numerical computation tools or software for approximation. This equation might not have an analytical solution in general terms but can be approximated using numerical methods to eventually determine the defocus distance Δ_{D_1}. Once the defocus distance Δ_{D_1} is known, the corresponding PSF can be established. In practical WSI scenarios, beyond considering unavoidable factors such as aberrations and noise, the layered depth-of-field model needs to take into account the PSF corresponding to N imaging layers. Hence, restoring an in-focus image analytically remains a challenge.

In summary, unlike the one-shot method, the dual-shot approach features a unified transmission model. Specifically, in the ideal case where only a single-layer DoF model is considered, the virtual autofocus method for a single image is essentially a blind deconvolution problem. That is, given a defocus blurred image, the corresponding in-focus sharp image and the PSF are unknown variables. In contrast, for the dual-shot

virtual autofocus method, theoretical analysis reveals that it essentially becomes a non-blind deconvolution problem. Here, both the defocus image and its corresponding PSF are known, leaving the in-focus image as the sole unknown variable, which simplifies the solution. Although in practical applications, when considering a layered DoF model, the imaging inversion process might not be solvable numerically, the additional exposure image can still provide prior knowledge, enhancing the quality of in-focus image restoration by neural networks.

$$E\left(\Delta_{D_1}, \Delta_{D_2}, f\right) = \left\|g_1 - \left(h_{\Delta_{D_1}} * f\right)\right\|^2 + \left\|g_2 - \left(h_{\Delta_{D_2}} * f\right)\right\|^2 \tag{A.2.7}$$

We know that $\Delta_{D_2} = \Delta_{D_1} + \Delta$, and we find (Δ_{D_1}, f) using:

$$\frac{\partial E}{\partial \Delta_{D_1}} = 0 \text{ and } \frac{\partial E}{\partial f} = 0 \tag{A.2.8}$$

where J_0

$$J_0(x) = \sum_{m=0}^{\infty} \frac{(-1)^m}{m!\Gamma(m+1)} \left(\frac{x}{2}\right)^{2m} \tag{A.2.9}$$

$$\Gamma(m+1) = m! \tag{A.2.10}$$

A.3 WSI Workflow

In this subsection, we exhibit the WSI workflow of our method, as shown in Fig. A.2.

A.3.1 WSI Workflow of Focus Prediction Autofocusing

- **Pre-processing**: *We design a pre-processing strategy to make the shot position guided by a course focus map.* According to the observation of [24, 40], there is an uneven distribution of all tiles' focus points along the optical axis. Besides, the slide is unavoidable tilt and misplacement [92]. An efficient solution is to create a course focus map by only several points for the above issues. Then we will find how the slide tilt and perform one-shot autofocusing for each tile along the gradient of the focus map. In our method, the range of defocus distance is from -10μm to +10μm. The average thickness of the pathological tissue is usually 5μm. Thus, the range of our method is adequate to cover most of the defocus fluctuations. If the focus point is out of the fine focus range, increasing the number of course focus points is more effective than enlarging the fine focus range. Therefore, we can utilize a simple course focus map to know the slide tilt and perform autofocusing for each tile guided by the course focus map.

- **Scanning and Defocus Shooting**: *We utilize the course focus map after pre-processing to perform scanning and shooting for the whole slide (all tiles).* Unlike the conventional autofocusing method (creating z-stack for the tile), ours only scans the slide in one pass along the course focus map surface. In this scenario, the defocus distance ΔD is the distance from the focus point of each tile to the in-focus shooting position. The topography variation of tissue samples is responsible for different ΔD. Thus, when the whole slide is scanned and shot tile-by-tile, we capture three kinds of tiles: in-focus tiles, positive defocus tiles, and negative defocus tiles. Besides, the performance of tiles with small ΔD need to be guaranteed, because the probability density of focus points for all

tiles is much more significant in the center of the sample thickness.

- **Network Processing**: *We develop a learning-based strategy for autofocusing via deep cascade networks.* The input is a defocus tile and the output is the predicted defocus distance. Then we integrate all the defocus distances as a final focus map of the tissue sample. We will elaborate the details of our networks in the next subsection.

- **Re-scanning and In-focus Shooting**: *We perform re-scanning and in-focus shooting by the focus map surveying.* Guided by the focus map, we can determine the defocus distance of each tile and adjust the position of the objective lens for in-focus imaging. Then we obtain all in-focus tiles after tile-by-tile scanning and shooting.

- **Stitching and Showing**: *The whole slide image is obtained by stitching all tiles and shown on the screen.* Then the whole slide image can be processed for downstream image analysis tasks.

A.3.2 WSI Workflow of Dual-shot Autofocusing

The complete workflow of our method bears similarity to that described in Appendix A.3.1, with the main difference being the replacement of the focus map creation and the second scanning exposure with the dual-shot virtual autofocusing network designed in our study. The detailed steps are as follows: (1) Estimate the initial focal length for all tiles of the pathological slide; (2) Conduct a full slide scan where, for each tile, two exposures are made at different focal lengths; (3) The two images are processed through DAFNet to retrieve the in-focus image.

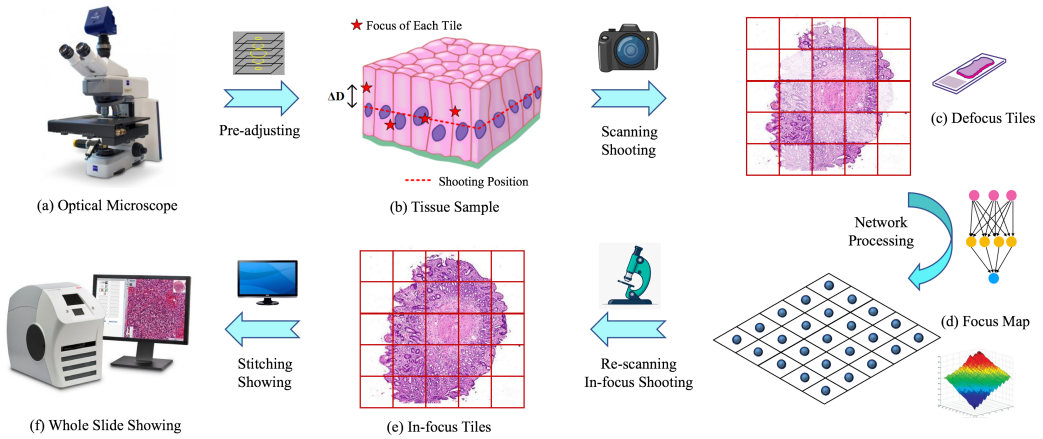

Figure A.2: The proposed workflow of WSI: (1) Pre-adjusting; (2) Scanning and Defocus Shooting; (3) Network Processing; (4) Re-scanning and In-focus Shooting; (5) Stitching and Showing.

A.3.3 WSI Workflow of One-shot Autofocusing

The comprehensive workflow of our method is akin to that described in Appendix A.3.1. The only modification required is the substitution of the focus map construction and the secondary scan exposure with the one-shot virtual autofocusing network designed in our approach. Specifically: firstly, a preliminary in-focus procedure is conducted on the central tile of the pathological sample to estimate the entire slide's focus plane position. Subsequently, the WSI system conducts x-y directional scanning, capturing all the defocus tiles, with the capture location set as the in-focus imaging position for the entire slide. Finally, the defocus tiles are fed into the neural network, and through algorithmic processing, restored in-focus tiles are derived. In essence, our method replaces the traditional z-stack focusing technique, substantially boosting the efficiency of procuring pathological slices.

Bibliography

[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.

[2] E. Abels, L. Pantanowitz, F. Aeffner, M. D. Zarella, J. vd Laak, M. M. Bui, V. N. Vemuri, A. V. Parwani, J. Gibbs, E. Agosto-Arroyo, et al. Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the digital pathology association. *The Journal of pathology*, 2019.

[3] D. A. Agard. Optical sectioning microscopy: cellular architecture in three dimensions. *Annual review of biophysics and bioengineering*, 13(1):191–219, 1984.

[4] M. Aittala and F. Durand. Burst image deblurring using permutation invariant convolutional neural networks. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision – ECCV 2018*, pages 748–764, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01237-3.

[5] Y. Bai, G. Cheung, X. Liu, and W. Gao. Graph-based blind image deblurring from a single photograph. *IEEE Transactions on Image Processing*, 28(3):1404–1418, 2018.

[6] A. G. Berman, W. R. Orchard, M. Gehrung, and F. Markowetz. Pathml: a unified framework for whole-slide image analysis with deep learning. *medRxiv*, pages 2021–07, 2021.

[7] 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. *Journal of Biophotonics*, 13(12):e202000227, 2020.

[8] M. Born and E. Wolf. *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light*. Elsevier, 2013.

[9] M. E. Bravo-Zanoguera, C. A. Laris, L. K. Nguyen, M. Oliva, and J. H. Price. Publisher's note: Dynamic autofocus for continuous-scanning time-delay-and-integration image acquisition in automated microscopy. *Journal of Biomedical Optics*, 12(3):39802–39900, 2007.

[10] A. Cable, J. Wollenzin, R. Johnstone, K. Gossage, J. S. Brooker, J. Mills, J. Jiang, and D. Hillmann. Microscopy system with auto-focus adjustment by low-coherence interferometry, Jan. 16 2018. US Patent 9,869,852.

[11] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019.

[12] C. Chen, M. Y. Lu, D. F. Williamson, T. Y. Chen, A. J. Schaumberg, and F. Mahmood. Fast and scalable search of whole-slide images via self-supervised deep learning. *Nature Biomedical Engineering*, 6(12):1420–1434, 2022.

[13] A. Cruz-Roa, H. Gilmore, A. Basavanhally, M. Feldman, S. Ganesan, N. N. Shih, J. Tomaszewski, F. A. González, and A. Madabhushi. Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach for quantifying tumor extent. *Scientific reports*, 7(1):46450, 2017.

[14] S. Cuenat, L. Andréoli, A. N. André, P. Sandoz, G. J. Laurent, R. Couturier, and M. Jacquot. Fast autofocusing using tiny transformer networks for digital holographic microscopy. *Optics Express*, 30(14):24730–24746, 2022.

[15] T. R. Dastidar and R. Ethirajan. Whole slide imaging system using deep learning-based automated focusing. *Biomedical Optics Express*, 11(1):480–491, 2020.

[16] N. Dimitriou, O. Arandjelović, and P. D. Caie. Deep learning for whole slide image analysis: an overview. *Frontiers in medicine*, 6:264, 2019.

[17] K. Ding, M. Zhou, H. Wang, O. Gevaert, D. Metaxas, and S. Zhang. A large-scale synthetic pathological dataset for deep learning-enabled segmentation of breast cancer. *Scientific Data*, 10(1):231, 2023.

[18] Y. Gan, Z. Ye, Y. Han, Y. Ma, C. Li, Q. Liu, W. Liu, C. Kuang, and X. Liu. Single-shot autofocusing in light sheet fluorescence microscopy with multiplexed structured illumination and deep learning. *Optics and Lasers in Engineering*, 168:107663, 2023.

[19] F. Ghaznavi, A. Evans, A. Madabhushi, and M. Feldman. Digital imaging in pathology: whole-slide imaging and beyond. *Annual Review of Pathology: Mechanisms of Disease*, 8:331–359, 2013.

[20] S. S. Goilkar and D. M. Yadav. Implementation of blind and non-blind deconvolution for restoration of defocused image. In *2021 International Conference on Emerging Smart Computing and Informatics (ESCI)*, pages 560–563. IEEE, 2021.

[21] J. W. Goodman. *Introduction to Fourier optics*. Roberts and Company Publishers, 2005.

[22] 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. *Optics Letters*, 45(1):260–263, 2020.

[23] K. Guo, J. Liao, Z. Bian, X. Heng, and G. Zheng. Instantscope: a low-cost whole slide imaging system with instant focal plane detection. *Biomedical Optics Express*, 6(9):3210–3216, 2015.

[24] M. Hart, R. H. Barkhouser, M. Carr, M. Golebiowski, J. E. Gunn, S. C. Hope, and S. A. Smee. Focal plane alignment and detector characterization for the subaru prime focus spectrograph. In *High Energy, Optical, and Infrared Detectors for Astronomy VI*, volume 9154, page 91540V. International Society for Optics and Photonics, 2014.

[25] C. Higgins. Applications and challenges of digital pathology and whole slide imaging. *Biotechnic & Histochemistry*, 90(5):341–347, 2015.

[26] M. S. Hosseini, J. A. Brawley-Hayes, Y. Zhang, L. Chan, K. N. Plataniotis, and

S. Damaskinos. Focus quality assessment of high-throughput whole slide imaging in digital pathology. *IEEE Transactions on Medical Imaging*, 39(1):62–74, 2019.

[27] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[28] L. Huang, T. Liu, X. Yang, Y. Luo, Y. Rivenson, and A. Ozcan. Holographic image reconstruction with phase recovery and autofocusing using recurrent neural networks. *ACS Photonics*, 8(6):1763–1774, 2021.

[29] B. Hulsken. Scanning imaging system with a novel imaging sensor with gaps for electronic circuitry, Oct. 2 2018. US Patent 10,091,445.

[30] 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. `https://figshare.com/articles/Data_set_for_Auto-focusing/5936881`.

[31] 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. *arXiv preprint arXiv:1905.03371*, 2019.

[32] M. Khened, A. Kori, H. Rajkumar, G. Krishnamurthi, and B. Srinivasan. A generalized deep learning framework for whole-slide image segmentation and analysis. *Scientific reports*, 11(1):11579, 2021.

[33] B. Kim and T. Naemura. Blind deconvolution of 3d fluorescence microscopy using depth-variant asymmetric psf. *Microscopy Research and Technique*, 79(6): 480–494, 2016.

[34] T. Kohlberger, Y. Liu, M. Moran, T. Brown, C. H. Mermel, J. D. Hipp, M. C. Stumpe, et al. Whole-slide image focus quality: Automatic assessment and impact on ai cancer detection. *arXiv preprint arXiv:1901.04619*, 2019.

[35] D. Krishnan, T. Tay, and R. Fergus. Blind deconvolution using a normalized sparsity measure. In *CVPR 2011*, pages 233–240. IEEE, 2011.

[36] P. Langehanenberg, G. von Bally, and B. Kemper. Autofocusing in digital holographic microscopy. *3D Research*, 2(1):4, 2011.

[37] W. Lee, H. S. Nam, Y. G. Kim, Y. J. Kim, J. H. Lee, and H. Yoo. Robust autofocusing for scanning electron microscopy based on a dual deep learning network. *Scientific reports*, 11(1):20933, 2021.

[38] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman. Understanding blind deconvolution algorithms. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2354–2367, 2011.

[39] C. Li, A. Moatti, X. Zhang, H. T. Ghashghaei, and A. Greenbaum. Deep learning-based autofocus method enhances image quality in light-sheet fluorescence microscopy. *Biomedical Optics Express*, 12(8):5214–5226, 2021.

[40] Q. Li, X. Liu, J. Jiang, C. Guo, X. Ji, and X. Wu. Rapid whole slide imaging via dual-shot deep autofocusing. *IEEE Transactions on Computational Imaging*, pages 1–1, 2020. doi: 10.1109/TCI.2020.3046189.

[41] 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. *Biomedical optics express*, 7(11):4763–4768, 2016.

[42] 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.

[43] 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. *Journal of biomedical optics*, 23(6):066503–066503, 2018.

[44] J. Liao, Z. Wang, Z. Zhang, Z. Bian, K. Guo, A. Nambiar, Y. Jiang, S. Jiang, J. Zhong, M. Choma, et al. Dual light-emitting diode-based multichannel microscopy for whole-slide multiplane, multispectral and phase imaging. *Journal of biophotonics*, 11(2):e201700075, 2018.

[45] J. Lightley, F. Görlitz, S. Kumar, R. Kalita, A. Kolbeinsson, E. Garcia, Y. Alexandrov, V. Bousgouni, R. Wysoczanski, P. Barnes, et al. Robust deep learning optical autofocus system applied to automated multiwell plate single molecule localization microscopy. *Journal of Microscopy*, 288(2):130–141, 2022.

[46] Y. Liron, Y. Paran, N. Zatorsky, B. Geiger, and Z. Kam. Laser autofocusing system for high-resolution cell biological imaging. *Journal of microscopy*, 221 (2):145–151, 2006.

[47] C.-S. Liu and S.-H. Jiang. A novel laser displacement sensor with improved robustness toward geometrical fluctuations of the laser beam. *Measurement Science and Technology*, 24(10):105101, 2013.

[48] C.-S. Liu and S.-H. Jiang. Design and experimental validation of novel enhanced-performance autofocusing microscope. *Applied Physics B*, 117:1161–1171, 2014.

[49] C.-S. Liu and S.-H. Jiang. Precise autofocusing microscope with rapid response. *Optics and Lasers in Engineering*, 66:294–300, 2015.

[50] C.-S. Liu, Z.-Y. Wang, and Y.-C. Chang. Design and characterization of high-performance autofocusing microscope with zoom in/out functions. *Applied Physics B*, 121:69–80, 2015.

[51] L. B. Lucy. An iterative technique for the rectification of observed distributions. *Astronomical Journal, Vol. 79, p. 745 (1974)*, 79:745, 1974.

[52] Y. Luo, L. Huang, Y. Rivenson, and A. Ozcan. Single-shot autofocusing of microscopy images using deep learning. *arXiv preprint arXiv:2003.09585*, 2020.

[53] 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.

[54] R. R. McKay, V. A. Baxi, and M. C. Montalto. The accuracy of dynamic predictive autofocusing for whole slide imaging. *Journal of pathology informatics*, 2, 2011.

[55] J. G. McNally, T. Karpova, J. Cooper, and J. A. Conchello. Three-dimensional imaging by deconvolution microscopy. *Methods*, 19(3):373–385, 1999.

[56] 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. *Journal of pathology informatics*, 2, 2011.

[57] M. Montoya, M. J. Lopera, A. Gómez-Ramírez, C. Buitrago-Duque, A. Pabón-Vidal, J. Herrera-Ramirez, J. Garcia-Sucerquia, and C. Trujillo. Focusnet: An

autofocusing learning-based model for digital lensless holographic microscopy. *Optics and Lasers in Engineering*, 165:107546, 2023.

[58] E. Nehme, D. Freedman, R. Gordon, B. Ferdman, L. E. Weiss, O. Alalouf, T. Naor, R. Orange, T. Michaeli, and Y. Shechtman. Deepstorm3d: dense 3d localization microscopy and psf design by deep learning. *Nature Methods*, 17(7): 734–740, 2020.

[59] R. J. Noll. Zernike polynomials and atmospheric turbulence. *JOsA*, 66(3):207–211, 1976.

[60] J. Pan, D. Sun, H. Pfister, and M.-H. Yang. Blind image deblurring using dark channel prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1628–1636, 2016.

[61] L. Pantanowitz, P. N. Valenstein, A. J. Evans, K. J. Kaplan, J. D. Pfeifer, D. C. Wilbur, L. C. Collins, and T. J. Colgan. Review of the current state of whole slide imaging in pathology. *Journal of pathology informatics*, 2, 2011.

[62] L. Pantanowitz, J. H. Sinard, W. H. Henricks, L. A. Fatheree, A. B. Carter, L. Contis, B. A. Beckwith, A. J. Evans, A. Lal, and A. V. Parwani. Validating whole slide imaging for diagnostic purposes in pathology: guideline from the college of american pathologists pathology and laboratory quality center. *Archives of Pathology and Laboratory Medicine*, 137(12):1710–1722, 2013.

[63] S. Park, Y. Kim, and I. Moon. Fast automated quantitative phase reconstruction in digital holography with unsupervised deep learning. *Optics and Lasers in Engineering*, 167:107624, 2023.

[64] H. Pinkard, Z. Phillips, A. Babakhani, D. A. Fletcher, and L. Waller. Deep learning for single-shot autofocus microscopy. *Optica*, 6(6):794–797, 2019.

[65] J. H. Price. Autofocus system for scanning microscopy having a volume image formation, Aug. 3 1999. US Patent 5,932,872.

[66] T. Rai Dastidar. Automated focus distance estimation for digital microscopy using deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[67] G. Reinheimer. Arrangement for automatically focussing an optical instrument, Mar. 20 1973. US Patent 3,721,827.

[68] W. H. Richardson. Bayesian-based iterative method of image restoration. *JoSA*, 62(1):55–59, 1972.

[69] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

[70] M. Sabokrou, M. Fayyaz, M. Fathy, and R. Klette. Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. *IEEE Transactions on Image Processing*, 26(4):1992–2004, 2017.

[71] J. Schlemper, J. Caballero, J. V. Hajnal, A. N. Price, and D. Rueckert. A deep cascade of convolutional neural networks for dynamic mr image reconstruction. *IEEE transactions on Medical Imaging*, 37(2):491–503, 2017.

[72] C. A. Schneider, W. S. Rasband, and K. W. Eliceiri. Nih image to imagej: 25 years of image analysis. *Nature methods*, 9(7):671, 2012.

[73] C. Scofield. 212-d depth-of-field simulation for computer animation. In *Graphics Gems III (IBM Version)*, pages 36–38. Elsevier, 1992.

[74] A. Shajkofci and M. Liebling. Spatially-variant cnn-based point spread function estimation for blind deconvolution and depth estimation in optical microscopy. *IEEE Transactions on Image Processing*, 29:5848–5861, 2020.

[75] Y. Shechtman, S. J. Sahl, A. S. Backer, and W. Moerner. Optimal point spread function design for 3d imaging. *Physical review letters*, 113(13):133902, 2014.

[76] C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.

[77] L. Silvestri, M. C. Muellenbroich, I. Costantini, A. P. Di Giovanna, L. Sacconi, and F. S. Pavone. Rapid: Real-time image-based autofocus for all wide-field optical microscopy systems. *BioRxiv*, page 170555, 2017.

[78] M. Subbarao and J.-K. Tyan. Selecting the optimal focus measure for autofocusing and depth-from-focus. *IEEE transactions on pattern analysis and machine intelligence*, 20(8):864–870, 1998.

[79] Y. Sun, S. Duthaler, and B. J. Nelson. Autofocusing algorithm selection in computer microscopy. In *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on*, pages 70–76. Citeseer, 2005.

[80] J. Tang, J. Wu, J. Zhang, Z. Ren, J. Di, and J. Zhao. Single-shot diffraction

autofocusing: Distance prediction via an untrained physics-enhanced network. *IEEE Photonics Journal*, 14(1):1–6, 2021.

[81] J. P. Vink, B. Hulsken, M. Wolters, M. B. Van Leeuwen, and S. H. Shand. System for generating a synthetic 2d image with an enhanced depth of field of a biological sample, Apr. 14 2020. US Patent 10,623,627.

[82] T. Virág, A. László, B. Molnár, A. Tagscherer, and V. S. Varga. Focusing method for the high-speed digitalisation of microscope slides and slide displacing device, focusing optics, and optical rangefinder, Feb. 16 2010. US Patent 7,663,078.

[83] J. Wei and T. Hellmuth. Optical coherence tomography assisted ophthalmologic surgical microscope, Feb. 20 1996. US Patent 5,493,109.

[84] R. S. Weinstein, A. R. Graham, L. C. Richter, G. P. Barker, E. A. Krupinski, A. M. Lopez, K. A. Erps, A. K. Bhattacharyya, Y. Yagi, and J. R. Gilbertson. Overview of telepathology, virtual microscopy, and whole slide imaging: prospects for the future. *Human pathology*, 40(8):1057–1069, 2009.

[85] Y. Wen, B. Sheng, P. Li, W. Lin, and D. D. Feng. Deep color guided coarse-to-fine convolutional network cascade for depth image super-resolution. *IEEE Transactions on Image Processing*, 28(2):994–1006, 2018.

[86] Y. Wu, V. Boominathan, H. Chen, A. Sankaranarayanan, and A. Veeraraghavan. Phasecam3d—learning phase masks for passive single view depth estimation. In *2019 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE, 2019.

[87] 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. *Nature methods*, 16(12):1323–1331, 2019.

[88] Y. Xu, X. Wang, C. Zhai, J. Wang, Q. Zeng, Y. Yang, and H. Yu. A single-shot autofocus approach for surface plasmon resonance microscopy. *Analytical Chemistry*, 93(4):2433–2439, 2021.

[89] J. Yan, P. DiMeo, L. Sun, and X. Du. Lstm-based model predictive control of piezoelectric motion stages for high-speed autofocus. *IEEE Transactions on Industrial Electronics*, 70(6):6209–6218, 2022.

[90] 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.

[91] M. D. Zarella, D. Bowman, F. Aeffner, N. Farahani, A. Xthona, S. F. Absar, A. Parwani, M. Bui, and D. J. Hartman. A practical guide to whole slide imaging: a white paper from the digital pathology association. *Archives of pathology & laboratory medicine*, 143(2):222–234, 2018.

[92] ZEISS. A Basic and Quick guide to Axio Scan.Z1. `https://hcbi.fas.harvard.edu/files/hcbidoug/files/axio_scan.z1_application_guide.pdf`, 2014.

[93] X. Zhang, Z. Dong, H. Wang, X. Sha, W. Wang, X. Su, Z. Hu, and S. Yang. 3d positioning and autofocus of the particle field based on the depth-from-defocus method and the deep networks. *Machine Learning: Science and Technology*, 2023.

[94] X. Zhou, R. Molina, Y. Ma, T. Wang, and D. Ni. Parameter-free gaussian psf model for extended depth of field in brightfield microscopy. *IEEE Transactions on Image Processing*, 29:3227–3238, 2019.

[95] Y. Zou, G. J. Crandall, and A. Olson. Real-time focusing in line scan imaging, Dec. 12 2017. US Patent 9,841,590.