Image Quality Assessment and Refocusing with Applications in Whole Slide Imaging

by

Zhongling Wang

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2024

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Jie Liang Professor, School of Engineering Science, Simon Fraser University
Supervisor(s):	Zhou Wang Professor, Dept. of Electrical and Computer Engineering, University of Waterloo
Internal Member:	Krzysztof Czarnecki Professor, Dept. of Electrical and Computer Engineering, University of Waterloo
Internal Member:	Oleg Michailovich Associate Professor, Dept. of Electrical and Computer Engineering, University of Waterloo
Internal-External Member:	Giang Tran Assistant Professor, Dept. of Applied Mathematics, University of Waterloo

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In the rapidly evolving field of general digital imaging and whole slide imaging, Image Quality Assessment (IQA) plays a crucial role in determining the perceptual quality of images and guiding image restoration. State-of-the-art IQA models are computationally expensive due to the use of complex deep learning architectures. The high computational cost poses a significant challenge in high-throughput Whole Slide Image (WSI) scanning platforms, which is both time-sensitive and power-limited. Moreover, most IQA models, while varied in design, often exhibit biases towards specific types of image content or distortions, a consequence of their underlying design principles or training data. To improve the quality of WSIs, we need to address the defocus problem, which is the most common distortion for a WSI. The transparency and uneven surface of tissue samples further complicate the restoration process for methods that lack an understanding of the 3D tissue radiance. These issues emphasize the limitations and challenges faced by existing IQA and restoration models. This thesis proposes three novel and flexible approaches to mitigate these problems.

Addressing the efficiency concerns in whole slide imaging, this thesis presents a highly efficient model for Focus Quality Assessment (FQA). Among the distortions that degrade the quality of digital slides, out-of-focus blur is the most common one. Different from photographic images, WSIs have much bigger dimensions, making most deep-learning based FQA models computationally infeasible. Based on prior knowledge of the WSI and its imaging process, we developed a lightweight model named FocusLiteNN that is $10,000$ times more efficient than SOTA deep learning-based ones without compromising accuracy. Furthermore, we introduce the first open-source, expert annotated FQA dataset TCGA@Focus, offering a comprehensive platform for developing and evaluating new FQA models.

However, most FQA models, or IQA models in general, often exhibit biases towards specific types of image content or distortions due to their different design principles or training data. This poses a challenge for users when choosing the best quality assessment model for their needs. A practical approach is to fuse the results of multiple existing IQA models into a more robust one. Following this idea, we developed a novel framework for

IQA score fusion that is able to select the best combination of models according to the uncertainty in each image and the overall uncertainty of each model. This requires the model to be equipped with both fine-grained uncertainty analysis at the content level and coarse-grained uncertainty analysis at the model level, respectively. Existing models either lack content-level uncertainty estimation or have limited generalizability due to supervised training. Our method employs an unsupervised approach using deep Maximum a Posteriori (MAP) estimation, which can be trained on a combination of multiple datasets without the need for Mean Opinion Score (MOS). This greatly improves the generalizability of the model.

The above two works address different problems in quality assessment. In practice, detected bad-quality images are either rejected or recollected. In digital pathology, recollecting the biosample causes additional suffering for the patient. Consequently, defocus restoration is a possible solution. Deblurring assumes that there exists a sharp image in which all pixels are in-focus, which is commonly referred to as a All-In-Focus (AIF) image. Although this assumption is true for natural images, it might not hold for WSIs due to its transparency, uneven surface and the microscope's shallow Depth of Field (DOF). Since the target does not exist, WSI deblurring becomes an undefined task. We propose an alternative approach to address the defocus problem, which is virtual refocusing. It aims to simulate and surpass the traditional experience of one continuously adjusting the focus of a microscope, allowing for a comprehensive examination of tissue structures at varying depths without the need for physical slide presence. By implicitly learning a continuous 3D radiance representation from the sparse inputs, the proposed model can refocus each pixel to any focus plane according to a focus map. As far as we know, this is the first work on WSI virtual refocusing.

This thesis makes significant contributions to IQA and image restoration with applications in WSI. The introduction of the FocusLiteNN model boosts computational efficiency while the score fusion model addresses the bias issue. Additionally, the virtual refocusing model extends these improvements by tackling the defocus problem in WSI through precise adjustment of focus on a per-pixel basis.

Acknowledgments

Completing this Ph.D. has been an incredible journey, and I wouldn't be here without the support of some amazing people.

Professor Zhou Wang, my advisor, deserves my sincerest thanks. He patiently taught me the fundamentals of good research, from critical thinking to presenting my work clearly. More importantly, he encouraged me to follow my own research interests, even if they were unexplored and challenging. I feel incredibly lucky to have him as my mentor.

I'm also deeply grateful to the entire IVC team for their constant help and friendship. Zhengfang Duanmu and Wentao Liu, your guidance as senior members was invaluable, both in the lab and in life. Wenbo Yang, I owe you a huge thanks for collaborating with me on many projects. I would also like to thank Shahrukh Athar, Zhuoran Li, Jinghan Zhou, Sheyang Tang, Xiaoyu Xu, Armin Shafiee Sarvestani, Mahdi Naseri, Mahzar Eisapour, Jiayan Qiu, Wei Zhou, Jiebin Yan, Yu Wang, Delu Zeng, Qi Liu, Honglei Su, Jia Yan, Diqi Chen for the insightful discussions and help.

Finally, to my family: words can't express my gratitude for your endless love and support. Mom and Dad (Xiaofang Yan and Geyu Wang), thank you for always believing in me and providing a safe harbor when I needed it most. And Fangyan Sun, my amazing girlfriend, you've been my rock throughout this challenging journey. Thank you for always being there to pick me up when I fell. Your cheerful laugh has cured me countless times and I wouldn't be here without you.

Dedication

This is dedicated to my family.

Table of Contents

Examining Committee Membership ii

Author's Declaration iii

Abstract iv

Acknowledgments vi

Dedication vii

List of Figures xii

List of Tables xvii

List of Abbreviations xxi

1 Introduction 1

1.1 Motivation 1

1.2 Objectives 5

1.3 Contributions 10

1.4 Thesis Outline 12

2 Background and Literature Review **13**

2.1 Focus Quality Assessment for Whole Slide Image 13

2.1.1 Focus Quality Assessment Models 13

2.1.2 Focus Quality Assessment Datasets 25

2.1.3 Image Sharpness Assessment Models 30

2.2 Image Quality Assessment Score Fusion 38

2.2.1 Taxonomy of IQA Models 38

2.2.2 Empirical IQA Score Fusion 41

2.2.3 Supervised learning-Based Score Fusion 43

2.2.4 Rank Fusion 44

2.2.5 Pairwise Ranking Guided NR-IQA Methods 46

2.3 Whole Slide Image Defocus Restoration and Synthesis 47

2.3.1 Whole Slide Image Deblur 47

2.3.2 Whole Slide Image Focus Interpolation 48

2.3.3 Whole Slide Image Defocus Synthesis 50

3 High-Efficiency Focus Quality Assessment for Whole Slide Image **51**

3.1 Introduction 51

3.2 FocusLiteNN Model 60

3.2.1 Difference in Natual Image and WSI 60

3.2.2 Assumptions 63

3.2.3 Model Design and Analysis 68

3.3 TCGA@Focus Dataset 79

3.4 Experiments 82

3.4.1 Implementation Details . . . 82
3.4.2 Performance Evaluation . . . 85
3.4.3 Ablation Study . . . 87
3.4.4 Computational Complexity Analysis . . . 87
3.4.5 Heat Map Visualization . . . 88

4 Unsupervised IQA Score Fusion by Deep Maximum a Posteriori Estimation 101

4.1 Introduction . . . 101
4.2 Proposed Framework . . . 103
4.2.1 Observation Model . . . 103
4.2.2 MAP Formulation and Optimization . . . 105
4.2.3 Amortized Inference . . . 107
4.2.4 Rank Fusion . . . 107
4.3 Experiments . . . 108
4.3.1 Implementation and Experimental Details . . . 108
4.3.2 Evaluation Results . . . 109

5 Whole Slide Image Virtual Refocusing 122

5.1 Introduction . . . 122
5.2 Method . . . 133
5.2.1 General Overview . . . 134
5.2.2 Implicit 3D Radiance Field Reconstruction . . . 136
5.2.3 Refocusing Through Focal Stack Cross-Attention Pooling . . . 139
5.2.4 Whole Slide Image Perceptual Distance Metric . . . 141

5.2.5 Overall Objective 146
5.3 Experiments 147
5.3.1 Implementation Details 147
5.3.2 Evaluation Results 148
5.3.3 Ablation Study 153

6 Conclusion and Future Work **162**

References **164**

List of Figures

2.1 The pipelines of major ways of collecting images and focus level labels for FQA. IF is the abbreviation of in-focus and OOF represents out-of-focus. Rounded rectangles filled with gray are the desired output. 26

2.2 Focus level distribution of the DeepFocus dataset [1]. 29

2.3 Sample images of the DeepFocus dataset [1] at different focus levels. 30

2.4 A sample of Z-stack images in the FocusPath dataset [2] 31

2.5 Sample images cropped from each of the slides in the FocusPath dataset [2] 32

2.6 A taxonomy for IQA models based on their design principles. 40

3.1 The tissue preparation and scanning pipeline for digital pathology. 52

3.2 The general structure of a WSI scanner. α is the half-angle of the cone of light entering the objective lens. 55

3.3 (a) Folding artifact on the WSI of bone [3]. The tissue's top layer appears to be in-focus, while the bottom layer is out-of-focus. The bottom layer still obstructs the upper layer's perceptual clarity even though the upper layer is in-focus. (b) Air bubble artifact caused by incorrect coverslipping. Because of the air bubble's distinct refractive index, light is diffracted differently from the other parts, resulting in a blurry image. 57

3.4 The distribution of the image gradient normalized by the average luminance. The distribution of sharp natural images is different from the one of sharp WSIs. 61

3.5 This figure [1]illustrates how semantic information affects the perceived sharpness of natural images. All parts of the original image are in focus. However, if individual local patches are isolated from the whole image, their perceived sharpness might be different from the whole image. Textured patches are sharp and patches containing smooth edges are blurry. The sharpness of smooth, constant, overexposed, or underexposed patches is undefined. . . . 65

3.6 This figure illustrates that the perceived sharpness of WSI is irrelevant to the semantic information. 67

3.7 (a) The systematic view of the chromatic aberration. Different focus distances result in different aberration patterns. (b) Example images captured at different focus distances. Note that the chromatic aberration is different in these two images: the upper one has some green/yellow color along the edges, while the lower one is red/blue. 69

3.8 This figure illustrates the influence of patch size on sharpness assessment. While both the yellow (235×235) and yellow-green (64×64) patches are in-focus, the yellow one looks sharper since it captures more biological structural information, such as the nucleus. While both the cyan (64×64) and green (235×235) patches are in the same level of out-of-focus, the cyan one looks sharper since the sensor noise is more pronounced. 70

3.9 How different pooling strategies affect the feature map of a Sobel kernel filtered edge image. While max pooling or min pooling alone can only capture one of the two edges. The max pooling + min pooling strategy captures both the increasing and decreasing edges in the original image. Average pooling produces a blurred feature map where the activation is less significant. In the figures, the value of the minimum pooling is inverted. . . 73

3.10 FocusLiteNN (1-kernel grayscale) filter visualization: (a) 2D spatial representation, (b) 3D spatial representation 76

3.11 FocusLiteNN (1-kernel grayscale) filter visualization: shifted Fast Fourier Transform (FFT) amplitude. 77

3.12 FocusLiteNN (1-kernel grayscale) filter visualization: vertical and horizontal cross sections for FFT amplitudes. 78

3.13 The feature maps produced by convolving the image with the learned FocusLiteNN kernel and the Sobel filter. 79

3.14 (a) 3D spatial representation of the vertical Sobel filter, (b) 3D spatial representation of the Laplacian of Gaussian (LoG) filter (c) 3D FFT amplitude of the vertical Sobel filter, (d) 3D FFT amplitude of the LoG filter 89

3.15 Organ distribution of in-focus and out-of-focus images of the TCGA@Focus dataset. 90

3.16 In-focus and out-focus examples of the TCGA@Focus dataset. 91

3.17 The illustration of the circle of confusion when the focus distance does not match the in-focus distance. 92

3.18 Focus distance vs radius of the circle of confusion. It is easy to find that the radius is almost symmetric around the in-focus distance. 92

3.19 Scatter plots of absolute z-level versus predicted scores on the FocusPath dataset. 93

3.19 Scatter plots of absolute z-level versus predicted scores on the FocusPath dataset (continued). 94

3.20 Histogram of objective scores on the TCGA@Focus dataset. 95

3.20 Histogram of objective scores on the TCGA@Focus dataset (continued). . . 96

3.21 ROC curves of the testing models evaluated on the TCGA@Focus dataset. 97

3.22 Absolute heatmaps. A higher score indicates more blurriness. The predicted scores represent the z-levels in the FocusPath dataset. 98

3.22 Normalized heatmaps. A higher score indicates more blurriness. The predicted scores of each model are independently linearly normalized to the range 0 to 1. (continued). 99

3.23 Average processing time versus ROC-AUC, model size versus ROC-AUC and MMACs versus ROC-AUC on the TCGA@Focus Dataset. The x-axis of each figure is on a log scale. All models are running on an Intel i9-7920X @ 2.90GHz with 32 GB memory. ROC-AUC: Area under the receiver operating characteristic curve. MMACs: Million multiply-accumulate operations. . . 100

4.1 The diagram of the proposed framework fusing IQA scores $\mathbf{x_i} = \{x_i^j | j = 1, 2, ..., M\}$ of image I_i. $\hat{x_i^j} = f^j(z_i)$ represents the reconstructed score. $\mathcal{SN}_i^j$ is the abbrev. of $\mathcal{SN}(\xi_i^j, \omega_i^j, \alpha_i^j)$, which is the predicted conditional Skew Normal distribution $p(x_i^j | z_i)$. 104

4.2 The empirical distribution of IWSSIM [4], FSIM [5] and VSI [6] evaluated on the KADID-10K [7] and VCLFER [8] dataset. The dashed red curve is the mean value of the scores (shown as blue dots). The solid orange curves are the conditional distributions, each representing the density of scores given a MOS range. It is easy to find that the conditional distributions are skewed toward the higher score side. Also, the variance of the conditional distribution is related to the MOS. 114

4.3 The empirical distribution of the normalized rankings of IWSSIM [4] and FSIM [5] evaluated on the KADID-10K [7] dataset. 115

5.1 Illustration of three types of Out-of-Focus. The green area is the DOF. The tissue that lies within it will be in-focus, while those that lie outside of it will be out-of-focus [2]. 124

5.2 Overview of the proposed virtual refocusing network. The network consists of two major components: (1) a 3D radiance field module for implicitly learning a 3D tissue representation and (2) a refocus module that refocuses this 3D representation based on a 2D focus map. 133

5.3 Architecture of the 3D radiance field module and refocus module. The overall model design is based on a U-net architecture. Both modules consist of three stages, with skip connections linking the outputs of each stage in the 3D radiance filed module to the corresponding stage in the refocus module. 135

5.4 Network structure of the 3D radiance encoder block and refocus decoder block. Each encoder block incorporates intra-image and inter-image attention modules, followed by 3D convolutions. Each decoder block consists of an SFT block and a FACA Pooling layer. 137

5.5 Example images of the Kimia Path24 Dataset dataset. Image credit: [9]. . 146

5.6 Virtual refocusing example. The inputs are Target 6 and Target 11. The refocused images are generated by setting the target focus maps to uniform ones ranging from 1 to 16. 156

5.6 Continous refocusing example. The inputs are Target 9 and Target 10. The refocused images are generated by setting the target focus maps uniformly to 9.2, 9.4, 9.6, and 9.8, respectively. 158

5.7 Extreme refocusing example. The input is Target 16, which is the most out-of-focus image in the focal stack. Refocused images are generated using uniform target focus maps ranging from level 1 to 16. 159

5.8 Non-uniform refocusing example. The inputs are Target 1 and Target 8, both exhibiting partial out-of-focus blur due to physical artifacts on the slide (note the horizontal strip at the bottom). The refocused image, generated using the non-uniform focus map shown in the third image, is in focus across all spatial locations. 160

5.9 Impact of manipulating 3D radiance features on refocusing. The left and middle images are refocused using only the first layer of the 3D radiance features, resulting in both images being focused on the topmost tissue layer despite different target focal planes. The right image, refocused using all feature layers, is correctly focused at the target focal plane. 160

5.10 Attention maps of the third FSCA Pooling layer. The left and right attention maps are generated when the target focus plane is set to 1 and 7, respectively. FSCA Pooling selectively attends to the layers of the 3D radiance feature representation that are most relevant to the target focus plane. 161

List of Tables

2.1 The details of the two public available FQA datasets. 28

2.2 Whole Slide Image FQA and deblur methods that use synthetic defocus training data. 50

3.1 The details of the proposed TCGA@Focus dataset, compared with the only two public available FQA datasets. IF and OOF are the abbreviations for in-focus and out-of-focus, respectively. The number of equivalent patches is calculated as the number of 1024×1024 patches. 80

3.2 SRCC, PLCC, ROC-AUC, PR-AUC Performance of 16 NR-ISA Metrics on FocusPath Dataset and TCGA@Focus Dataset. The number of parameters, average processing time, and computational complexity are also reported. . 85

3.3 Statistical significance testing of FQA methods on the FocusPath dataset using prediction residuals. 1 means that the method is statistically better than the method in the column, 0 means that it is statistically worse, and - means that it is statistically indistinguishable. 86

3.4 A performance comparison of using fixed kernels in FocusLiteNN. 87

3.5 A performance comparison of regression models: Max&Min Pooling, Multi-layer Perceptron (MLP), Support Vector Regressor (SVR), RBFNet. The features used are the same for all regression models. The (inference) time is calculated relative to the Max&Min Pooling model. The feature extraction time is excluded from this test. ROC and PR are calculated based on binary classification. 87

3.6 A performance comparison of loss functions in training the FocusLiteNN (1-kernel) model: Pearson Linear Correlation Coefficient (PLCC), Concordance Correlation Coefficient (CCC), Mean Square Error (MSE), Mean Absolute Error (MAE), multi-class Cross Entropy and multi-class Ordinal Cross Entropy. ROC and PR are calculated based on binary classification. 88

4.1 Evaluation results in terms of Spearman Rank Correlation Coefficient (SRCC) of the proposed model, five other fusion methods and 16 individual IQA models used in the fusion. The average SRCC values are computed as weighted sums of individual ones, with weights determined by the number of images in each set. The top three best-performing models are shown in **bold** font while supervised methods are excluded. 116

4.2 Evaluation results in terms of PLCC of the proposed model, five other fusion methods and 16 individual IQA models used in the fusion. The average PLCC values are computed as weighted sums of individual ones, with weights determined by the number of images in each set. The top three best-performing models are shown in **bold** font while supervised methods are excluded. 117

4.3 Statistical significance testing of fusion methods using prediction residuals. Each entry is a codeword made up of ten symbols, with each symbol representing the test result for an IQA database. The order of the database is the same as in Table 4.1. 1 means that the method is statistically better than the method in the column on that particular dataset, 0 means that it is statistically worse, and - means that it is statistically indistinguishable. . 118

4.4 Evaluation results when "bad" IQA metrics are added to the list of models to be fused. 119

4.5 Evaluation results when IQA datasets with "bad" MOS are used as the training set. 119

4.6 Evaluation results when small IQA datasets are used as the training set. . 119

4.7 Evaluation results of FQA score fusion. We fused seven FQA methods on the FocusPath [2] dataset. The top three best-performing models are shown in **bold** font. 120

4.8 Inference speed comparison of different fusion methods. The time reported is for fusing a single image based on evaluations from 16 IQA models. Note that the inference speed of rank fusion methods is also influenced by the number of images in the dataset, which is set to 10,000 in this experiment. 121

5.1 Comparison of different types of microscopes and their abilities to handle the three types of out-of-focus. However, most methods are not directly applicable to WSIs in digital pathology due to practical constraints and the need for specialized, often expensive, equipment. 125

5.2 Comparison of different tasks related to WSI virtual refocusing. Natural image-based tasks assume scene non-transparency, making them unsuitable for WSI. The goal in WSI is to visualize 3D tissue structure. Generating a single AIF image, however, collapses this 3D information, hiding the desired structural details. 127

5.3 Refocusing performance with varying numbers of input images. Using more input images leads to better performance, as they capture richer 3D information about the radiance field. 151

5.4 Focus interpolation comparison results. The proposed model outperforms the other two WSI focus interpolation models. 151

5.5 A comparison of WSI deblurring results. The proposed model outperforms the other three deep learning-based debluring models. 152

5.6 Inference speed comparison of the refocus model and WSI interpolation methods. The time reported is for generating a 512×512 patch using one NVIDIA GTX 3090 GPU. 153

5.7 Alabtion study on the effectiveness of the 3D Radiance Module. It can be seen that using this module significantly increases the quality of the refocused images. 154

5.8 Alabtion study on the effectiveness of the Focal Stack Cross-Attension Pooling module. Compared to Max and Mean pooling, FSCA Pooling takes the 3D structure of the feature and the condition into account. It can be seen that using FSCA significantly increases performance. 154

5.9 Alabtion study on the effectiveness of the WSI-DISTS reconstruction loss. Compared to the original DISTS, WSI-DISTS is fine-tuned on WSI data, which makes it more accurate in measuring the distance between WSIs. . . 155

List of Abbreviations

AIF All-In-Focus v, xix, 3, 4, 8, 9, 126–129, 133, 140, 150

CCC Concordance Correlation Coefficient xviii, 88

CDSA Cancer Digital Slide Archive 27

CNN Convolutional Neural Network 5, 6, 9, 22, 24, 27, 29, 37, 58, 59, 63, 66, 71, 72, 74, 79, 140, 142

DCT Discrete Cosine Transform 16, 17, 35, 36

DOF Depth of Field v, xv, 2–4, 31, 47, 53–56, 62, 66, 123, 124

DoG Difference of Gaussian 5, 18, 33, 71

DR IQA Degraded-Reference IQA 39

DWT Discrete Wavelet Transform 17, 71

FFPE Formalin-Fixed Paraffin-Embedded 52, 129

FFT Fast Fourier Transform xiii, xiv, 71, 77, 78, 89

FIR Finite Impulse Response 22

FQA Focus Quality Assessment iv, xii, xvii, xix, 2–6, 8, 10, 12, 13, 19, 20, 22, 24–28, 30, 33, 37, 38, 50, 56–60, 63, 66, 68, 70–72, 74–76, 78–81, 112, 120, 134, 150, 162, 163

FR IQA Full-Reference IQA 20, 39, 131, 136, 141, 142, 144, 146, 150

H&E Hematoxylin and Eosin 21, 28, 29, 145

HVS Human Visual System 13, 18, 22, 26, 33, 36, 37, 39, 141, 142, 144

IHC Immunohistochemistry 21, 56, 145

IQA Image Quality Assessment iv, v, xii, xv, xix, 1–3, 5, 7, 8, 11, 12, 25, 37–44, 46, 58, 75, 76, 101–104, 110, 112, 121, 131, 162, 163

ISA Image Sharpness Assessment 6, 12, 21, 22, 30, 33, 34, 37, 58, 60, 62, 63, 66, 71, 76

KRCC Kendall Rank Correlation Coefficient 45

LoG Laplacian of Gaussian xiv, 5, 71, 77–79, 87, 89

MAE Mean Absolute Error xviii, 75, 88, 131, 136, 141, 142

MAP Maximum a Posteriori v, 3, 8, 11, 102, 103, 130, 162

MLE Maximum Likelihood Estimation 45

MLP Multi-layer Perceptron xvii, 36, 43, 73, 87, 109, 113, 121

MOS Mean Opinion Score v, 3, 7, 11, 41–44, 101, 102, 111, 144, 162

MSE Mean Square Error xviii, 9, 20, 24, 46, 75, 88, 131, 136, 141–143

NA Numerical Aperture 31, 53, 56, 62

NAS Neural Architecture Search 6

NR IQA No-Reference IQA 20, 37, 39

NSS Natural Scene Statistics 13, 22, 33, 36

PLCC Pearson Linear Correlation Coefficient xviii, 75, 76, 88, 110, 117

PSF Point Spread Function 4, 5, 8, 10, 22, 24, 27, 34, 47, 48, 50, 83, 129–131, 134, 136, 138, 163

ROI Region of Interest 54

RR IQA Reduced-Reference IQA 39

RRF Reciprocal Rank Fusion 44, 45

SGD Stochastic Gradient Descent 22

SNR Signal to Noise Ratio 64, 128

SRCC Spearman Rank Correlation Coefficient xviii, 45, 110, 116

SVM Support Vector Machine 13, 21, 72

SVR Support Vector Regressor xvii, 14, 34, 36, 43, 72, 87, 113

TCGA The Cancer Genome Atlas 27, 81

WSI Whole Slide Image iv, v, xii, xiii, xix, xx, 1–6, 8–12, 20–23, 25–28, 30, 31, 38, 47, 50, 52, 54–64, 66–68, 71, 79–82, 112, 122, 124–133, 136, 139, 142, 144, 145, 150, 151, 153, 155, 162, 163

Chapter 1

Introduction

1.1 Motivation

Efficient FQA

The field of digital pathology has been revolutionized by WSI, which allows slides to be examined remotely in digital format. However, this technique also introduces significant challenges, especially given the large data volumes these images generate. For example, the imaging process and post-processing are time-consuming and require a substantial amount of computational resources. In clinical practice, the WSI scanning is usually done overnight, and the scanned slides should be ready to be examined by pathologists the next morning. This tight schedule requires any image processing involved to be very efficient. Among the image-processing steps, IQA is a crucial one in the quality assurance process of digital pathology. It evaluates and filters out images with poor quality, which degrades the diagnostic accuracy. State-of-the-art IQA models are computationally expensive due to the use of complex deep learning architectures. The high computational cost poses a significant challenge in high-throughput WSI scanning platforms, which is both time-sensitive and power-limited. The motivation behind this work is to enhance the efficiency of IQA to meet the requirements of high-throughput WSI platforms without compromising accuracy. Among the distortions that degrade the quality of digital slides, out-of-focus blur

is the most common one. Most physical artifacts on the slides, such as air bubbles, dust, marker ink and tissue folds, will mislead the autofocus system. Since the DOF of the microscopy lens is very shallow compared to the tissue thickness, all these artifacts will lead to global or partial out-of-focus images. Addressing the efficiency concerns, this work presents a highly efficient model for FQA, which assesses the level of blur or out-of-focus level of a WSI. FQA can be considered as a special case of IQA that focuses on the defocus distortion.

IQA Score Fusion

However, most FQA models, or IQA models in general, often exhibit biases towards specific types of image content or distortions due to their different design principles or training data. This poses a challenge for users when choosing the best quality assessment model for their specific data, requirements, and applications. However, this job is nontrivial since the underlying assumptions of the model and the domain shift between the training and testing data are difficult to quantify. To address this issue, a straightforward idea is to develop a more powerful and general model that can cover all types of image contents and distortions. However, it would be difficult to account for every possible combination of image content and distortion, making it impractical to develop a single model that can handle all scenarios. A more practical approach is to fuse the results of multiple existing IQA models into a stronger one. Following this idea, we developed a novel framework for IQA score fusion that leverages the strengths and mitigates the weaknesses of individual models.

Existing fusion approaches can be categorized non-mutually exclusively into empirical, rank fusion methods, and supervised learning-based. Empirical models fuse a predetermined set of IQA models using a handcrafted formula. This approach significantly constrains its adaptability when introduced with new IQA models. Rank fusion methods operate in the discrete rank domain, where the range of all IQA models is mapped to the same uniform distribution. However, these methods are tied closely to the diversity of the ranking dataset, which can impede generalizability. Also, it can not handle the case when "bad" performing models are included in the fusion list. Supervised learning-based

methods are trained under the guidance of the MOS of a single subjective rated dataset. Such fusion methods are essentially refined versions of supervised learning-based IQA models since they are guided by the same ground truth, i.e., MOS. These models can not be trained on a combination of multiple datasets due to the MOS mismatch. Nevertheless, these black-box models often suffer from limited generalizability and lack of explainability. All the models mentioned above lack fine-grained uncertainty analysis at the content level, with some only offering coarse-grained uncertainty analysis at the model level or none at all. These limitations make the above-mentioned models less versatile and robust. What we need is a fusion method that is 1) flexible to incorporate any combination of models and be trained on any combination of datasets. 2) It should also be capable of rejecting bad-performing models as well as outlier content. 3) It should also have good generalizability and explainability. In this work, we try to achieve the above goals using an unsupervised learning-based framework based on MAP estimation. This is the first unsupervised IQA score fusion method as far as we know. This fusion framework also applies to FQA models. We demonstrate its flexibility and robustness on both natural images and WSIs.

WSI Virtual Refocusing

An efficient FQA model is essential in determining whether a digital slide's quality is good enough for diagnosis by a professional pathologist. But what happens if one slide does not meet the quality requirements? The answer is to collect the tissue and prepare the slide again for a new scan. This necessitates additional surgery to remove the tissue from the patient, leading to further discomfort for the patient and delaying the diagnostic process. In order to avoid redoing the whole process again, is there a way we can recover the information we need from the out-of-focus WSIs? From the perspective of post-processing, WSI deblurring is a possible solution. Deblurring assumes that there exists a sharp image in which all pixels are in-focus, which is commonly referred to as a AIF image. Although this assumption is true for natural images, it might not hold for WSIs.

The gap is arising from two different perspectives. 1) From the lens optic's perspective, the DOF of the microscopy lens is too shallow compared to the tissue thickness. Considering the surface of the tissue is usually noneven due to its biological nature, this means that

only the part of the tissue that lies in the DOF will be in focus. As a result, it is difficult to capture a AIF image in a WSI scanner. On the other hand, by reducing the diameter of the aperture, one can almost surely capture a AIF image using generic photography equipments. 2) From the slide's perspective, biological tissue slides are transparent and that is how a bright field microscope works. A transparent 3D object means that the inner structures of the tissue are visible and will be projected onto the image plane. However, most natural scenes are non-transparent. Although the atmosphere can be transparent, we do not care about the inner objects in it, if there are any, when doing natural image deblurring. This means that the captured image is actually a projection of a 2D manifold. The image formation process of a WSI can be simplified to convolving the 3D object with a set of 3D Point Spread Function (PSF)s, while the formation of natural images is convolving the 2D object (manifold) with a set of 2D PSFs. Since the tissue is entirely illuminated by the light source, the out-of-focus light passes through the deeper layers of the tissue that lie outside of the DOF and will interfere with the in-focus light. As a result, the captured image will be a superposition of in-focus and out-of-focus light, which makes the in-focus region blurry.

Due to the above two reasons, there does not exist such a AIF target for WSI deblurring. Even if it does, the AIF image will not be 100% sharp. This limitation makes deblurring an undefined task. We propose an alternative approach to address the defocus problem, which is virtual refocusing. Instead of finding a AIF image, virtual refocusing aims to simulate the most traditional experience of one continuously adjusting the focus knob of a microscope. Although there is no single focus plane in which all pixels are in focus, we can view each in-focus region by adjusting the focus continuously. Refocusing is the traditional practice before the emergence of digital pathology and WSI, but it requires the slide to be physically presence. In virtual refocusing, we rely on a single image taken at any focus level to recover the whole focus stack. We hypothesize that this process implicitly learns the 3D structure of the tissue. Consequently, giving images captured at multiple focus levels as input helps to enrich the implicit 3D representation learned by the process. As far as we know, this is the first work on WSI virtual refocusing.

Another major application of virtual refocusing is to synthesize realistic defocus distortions. Defocus synthesis has lots of applications in FQA, which can generate defocused

images with labels. Existing defocus synthesis methods are based on handcrafted blur kernels such as Gaussian, box kernel, and estimated PSF. These simple distortions ignore the 3D structure of the tissue and fail to generate realistic distortions.

To summarize, this thesis aims to address the need for an efficient FQA model, a versatile and robust IQA score fusion framework, and a WSI virtual refocusing model. In the next section, we introduce in detail the goals we would like to achieve based on these motivations.

1.2 Objectives

Efficient FQA

In the FQA literature, most knowledge-based models adopt handcrafted features, such as edges and high-frequency components. These features can usually be approximated by convolving the original image with a set of kernels. For example, commonly used edge detection kernels include Canny, Sobel, Prewitt, Scharr, Laplacian, Difference of Gaussian (DoG), LoG, etc. This demonstrates the possibility of using simple filters to extract sharpness-related features. Although relatively efficient, these handcrafted features may be suboptimal for the complex and diverse natural images. Recent advances in deep learning have demonstrated the power of Convolutional Neural Network (CNN) being able to extract task-specific features through learning from data. However, these complex designs increase the computational complexity, which makes most of the CNN architectures unsuitable for high throughput scanning platforms. In order to benefit from end-to-end training that produces more optimal solutions while being highly efficient at the same time, we aim to design a lightweight CNN-based FQA model. This requires us to explore a set of strong prior knowledge of WSIs and FQA and use them to guide the design of the network.

Compared to the diverse natural images and their complex distortions, we observe that WSIs have the following special characteristics: 1) WSIs are relatively uniform in feature scales, 2) the distortion process is well-controlled, and 3) the sharpness information is embedded in low-level features. More thorough justifications are provided in Sec 3.2.

Based on these observations, we argue that FQA for WSI is a simpler task compared to Image Sharpness Assessment (ISA) for natural images. Consequently, a multi-layer network is not necessary for FQA. On the contrary, we show that a single convolution kernel is sufficient to achieve a high accuracy that is on par with the state-of-the-art deep CNN models. Different from knowledge-based models that also depend on linear filters, the proposed model is end-to-end learned. This means that the network weight is adjusted to the specific need of the FQA task. Furthermore, the design of the model architecture is also based on the prior knowledge of the task. More specifically, according to the lens optics, we carefully design the structure of the convolution kernel based on the radius of the circle of confusion. We also select the appropriate nonlinear activation function according to the bandpass nature of the kernel. In addition to designing the architecture manually, Neural Architecture Search (NAS) offers an automated solution for finding the best-performing network within the search space. Despite this, the searched "optimal" result may not be simpler than the proposed model with a single convolutional kernel.

Another challenge for FQA is that all publicly available datasets only contain defocus distortion captured using the z-stack method. These slides are carefully selected to ensure high quality, and the defocus distortion is captured by moving the lens away from the optimal focus plane intentionally. They are also limited to slides prepared by one lab using the same protocol and scanned using one WSI scanner. Details are provided in Sec 3.3. This could mean that the dataset may not be diverse enough, which could result in a lower model generalization ability and biased evaluation accuracy when utilized for training and testing, respectively. To mitigate this problem, we proposed the first publicly available FQA dataset containing authentic defocus distortion. The images are extracted from slides in the TCGA [10] repository, which has very diverse content since they are collected and processed in labs around the world. The dataset contains $14,371$ patches of the size 1024×1024, with each patch annotated by experts.

IQA Score Fusion

While one FQA model may surpass another in specific situations, it is important to recognize that there is no universal model that outperforms all others in every aspect. This

is also true for IQA models in general. Each model exhibits biases towards specific types of image content or distortions due to their different design principles or training data. This poses a challenge for users when choosing the best quality assessment model for their specific applications. A practical and straightforward solution is to fuse the results of multiple IQA models into a stronger one. This fusion framework should be able to select the best combination of models according to each image and the overall performance of each model. This requires the model to be equipped with both fine-grained uncertainty analysis at the content level and coarse-grained uncertainty analysis at the model level, respectively. However, existing models in the literature lack explicit content-level uncertainty estimation, with some only offering model-level ones or none at all.

While not explicitly modeling content-level uncertainty, supervised learning-based methods may implicitly learn this from ground truth data, which is MOS. End-to-end learning-based methods have become popular recently due to their ability to find an optimal solution automatically. However, the generalization ability of these methods can be limited due to the domain gap between the training and testing data. These models can also suffer from overfitting if the training samples are not sufficient or diverse enough. One way to solve this problem is to collect more training data, which is usually achieved by combining multiple datasets. But this is very challenging for IQA datasets since the meaning of the ground truth, MOS, varies from dataset to dataset and can not be compared directly. This is caused by the use of different subjective experiment protocols, different groups of subjects, and different data processing methods. Is there any way we can bypass this problem of MOS mismatch? Our answer to this question is unsupervised learning. Without the need for ground truth, the model can be trained on any combination of datasets and has better generalization ability.

To summarize, our objective is to develop a IQA score fusion framework that is unsupervised-learning based, with both fine-grained uncertainty analysis at the content level and coarse-grained uncertainty analysis at the model level. A well-designed framework should harness the strengths and minimize the weaknesses of different models. This involves completely rejecting underperforming models and partially disregarding inaccurate predictions from strong models. It can also be trained on any combination of datasets, so its generalization ability is also good. Additional requirements include better explainability, a theoretically

sound formulation, and fast interference speed. In this work, we try to achieve the above goals using an unsupervised learning-based framework based on MAP estimation. This is the first unsupervised IQA score fusion method as far as we know. This fusion framework also applies to FQA models, which is a special case of IQA. We demonstrate its flexibility and robustness on both natural images and WSIs.

WSI Virtual Refocusing

The above two works address different problems in quality assessment, which is a crucial part of quality assurance in lots of applications. In most cases, detected bad-quality images are either rejected, recollected, or restored. For example, if a digital slide's quality does not meet the requirements for diagnosis, one needs to recollect the tissue and prepare the slide again for a new scan. To avoid redoing the whole process again, restoration, or WSI deblurring to be more specific, is a possible solution. Deblurring assumes that there exists a sharp image in which all pixels are in-focus, which is commonly referred to as a AIF image. Although this assumption is true for natural images, it might not hold for WSIs. Detailed reasons are described in Sec 1.1. Since the target for WSI deblurring does not exist, it becomes an undefined task. We propose an alternative approach to address the defocus problem, which is virtual refocusing. Instead of finding a AIF image, virtual refocusing aims to simulate the most traditional experience of one continuously adjusting the focus knob of a microscope. As far as we know, this is the first work on WSI virtual refocusing.

Different from physical refocusing, whose input is a physical slide, virtual refocusing relies on a single image taken at any focus level to recover images taken at different focus levels, which is referred to as the focus stack. Since each image is a result of convolving the 3D radiance field of the tissue with a set of 3D PSFs at different axial locations, an accurate estimation of the focus stack requires access to the 3D radiance field and the PSFs. However, the 3D radiance field is very difficult to capture without specific tissue processing and imaging equipment. The PSFs also need to be collected for each objective lens at all axial levels, which is a tedious process. To circumvent this problem, we develop an end-to-end virtual refocusing model that implicitly learns the 3D representation of the tissue without accessing the 3D structure or the PSFs.

Since one captured image does not contain much information about the 3D radiance field, using multiple captured images within a focal stack as input will help enrich the 3D information. Consequently, the proposed model should be able to leverage and fuse the information of the arbitrary number of images in a focal stack. The model also needs to learn to select the information within the implicit 3D representation that is related to the target focus plane and use it to generate the refocused image. This requires a focal stack-wise attention mechanism that attends to the information required by the target focus plane.

It would also be beneficial for the model to have the ability to refocus each pixel to a different target focus plane. This means that the target is a continuous focus map instead of a number indicating the axial location of the focus plane. In physical refocusing, we can only refocus the entire tissue to a constant focus plane. Due to the uneven tissue surface, it is common for not all regions to be in-focus simultaneously. If the model can refocus each pixel to a different focus plane, it is possible to generate an image that all regions are in focus at the same time. However, this AIF image will not be 100% sharp due to the transparent nature of the tissue. Detailed explanations are provided in Sec 1.1. Regardless, this is what we can achieve without explicitly modeling the 3D radiance field of the tissue.

Since the model is trained in a supervised manner, it is necessary to have a reconstruction loss function that measures the perceptual distance between the generated and the ground truth WSIs. Loss functions [11, 12] incorporating mid-level and high-level features have been shown to outperform those that only focus on low-level features such as MSE and SSIM [13]. These mid-level and high-level features are extracted from a pre-trained CNN, which is often trained on pristine natural images for the purpose of object classification. Since natural images are different from WSIs in serval perspectives as described in 3.2, these loss functions are suboptimal for measuring the distance between WSIs. For training the virtual refocusing model, what we need is a loss function tailored for WSIs. This can be achieved by replacing the feature extractor and fine-tuning the weights in the loss function. As far as we know, this is the first loss function tailored for WSIs.

Besides refocusing on the focus plane that gives the sharpest image, we are also interested in the images refocusing on out-of-focus planes. This corresponds to WSI defocus synthesis. Existing defocus synthesis methods are based on handcrafted blur kernels such

as Gaussian, box kernel, and estimated PSF. These simple distortions ignore the 3D structure of the tissue and the specifications of the optics. As a result, they often fail to generate realistic defocus distortions. As one of the built-in functions, the virtual refocusing model can be used to synthesize realistic homogeneously and in-homogeneously defocused images.

To summarize, the virtual refocusing model needs to learn the implicit 3D representation of the tissue without accessing the 3D radiance field and the PSFs. It also needs to accept an arbitrary number of images in a focal stack as input to enrich the 3D information. To generate a refocused image that all regions are in focus, the model needs to be conditioned on a continuous focus map instead of a number indicating the axial location of the focus plane. Finally, the training of the model requires a reconstruction loss function that is tailored for the specific needs of WSIs.

1.3 Contributions

In this section, we will summarize the main contributions of this thesis by reviewing the key contributions of each project.

Efficient FQA

FocusLiteNN is a lightweight FQA model designed to meet the efficiency requirements of high-throughput scanning platforms. Based on the prior knowledge of the uniformity of feature scales, the simplicity of distortion, the localization of sharpness information in low-level details, and the importance of color, the network consists of a single carefully designed convolutional kernel. The main contributions are

1. It is $10,000\times$ more efficient than the state-of-the-art networks while maintaining a comparable performance.
2. We proposed an expert annotated FQA dataset name TCGA@Focus, containing $14,371$ patches of the size 1024×1024. The images are very diverse since they are collected and processed in labs around the world. This is the first publicly available FQA dataset that features authentic defocus distortion.

IQA Score Fusion

For IQA score fusion, we develop an unsupervised learning-based fusion framework. The main contributions of our work include:

1. To the best of our knowledge, this is the first unsupervised learning-based score fusion method for IQA. It can be trained on any combination of datasets without the need of MOS.
2. We formalize the first observation model of IQA fusion and address the task using MAP estimation.
3. By building powerful coarse-grain and fine-grain uncertainty estimation modules, the proposed model harnesses the strengths and mitigates the weaknesses of each model. This means that it can completely reject underperforming models and partially disregard inaccurate predictions from strong models.
4. We show that rank fusion can be easily integrated into our general framework.

WSI Virtual Refocusing

The WSI virtual refocusing model aims to simulate and surpass the traditional experience of adjusting the focus of a microscope continuously. The contributions of the work are as follows:

1. As far as we know, it is the first WSI virtual refocusing model.
2. The model learns the implicit representation of the 3D radiance field based on a novel 3D consistency constraint. It also accepts an arbitrary number of images in a focal stack as input to enrich the 3D information.
3. Using a novel focal stack cross-attention module, the model selects the information related to the target focus map and uses it to generate the refocused image.

4. The model can refocus each pixel to a different focus plane in a continuous manner, which helps to generate an image that all regions are in focus at the same time.
5. It is the first WSI defocus synthesis model that features realistic distortions.
6. Using a novel image distance metric tailored for WSI as the reconstruction loss significantly improves the performance.

1.4 Thesis Outline

The structure of the following thesis is organized as follows:

- Chapter 2 provides a comprehensive background study and literature review of IQA score fusion, FQA for WSI, general ISA and WSI deblur.
- Chapter 3 describes the methodology, implementation, and experiment of the proposed highly efficient FQA model, named FocusLiteNN. It also includes the description of the proposed TCGA@Focus FQA dataset.
- Chapter 4 explores the unsupervised IQA score fusion framework, including its theoretical framework and experimental validations.
- Chapter 5 details the development of the WSI virtual refocusing model.
- Finally, Chapter 6 concludes the thesis with a summary and implications for future research.

Chapter 2

Background and Literature Review

2.1 Focus Quality Assessment for Whole Slide Image

2.1.1 Focus Quality Assessment Models

The FQA models can be divided into knowledge-based and data-driven ones based on the design philosophy behind the features used. These models rely on handcrafted features that are designed primarily based on the characteristics of Human Visual System (HVS), knowledge of out-of-focus distortion, and Natural Scene Statistics (NSS) of sharp images, etc. On the other hand, data-driven FQA models learn features from data through techniques such as deep learning.

Knowledge-based FQA

Knowledge-based FQA models depend on manually developed features that are mostly based on the characteristics of the HVS, understanding of out-of-focus distortion, and NSS of sharp (microscopic) images. The collected features are subsequently distilled into a boolean or scalar value that denotes the degree of blurriness, depending on whether the job is binary classification or regression. For classification tasks, this is often accomplished by employing either manual threshold or machine learning approaches such as Support

Vector Machine (SVM) [14], AdaBoost [15], decision tree, etc. For regression tasks, one can utilize either handcrafted linear/non-linear formulas or machine learning approaches such as SVR, linear regression, or logistic regression.

Local Statistics-based Features Common features being used include local patch statistics in either spatial or frequency domain, gradient and edge-related features, etc. Spatial statistics include local variance, mean value normalized local variance, autocorrelation, Shannon entropy, etc. To formalize the above features, we first introduce some notations. Without loss of generality, we assume the image to be evaluated is grayscale. Let the image patch denoted by $I \in \mathcal{R}^{H \times W}$, where H and W are the height and width of the image, respectively. $\mu = \frac{1}{HW} \sum_x \sum_y I(x, y)$ is the mean intensity of the patch, where (x, y) is the pixel coordinate. Local variance is defined as

$$F_{Var} = \frac{1}{HW} \sum_x \sum_y (I(x, y) - \mu)^2. \tag{2.1}$$

The variance can serve as a reliable measure of focusing sharpness since the variance of a well-focused image is generally higher than the one of a blurry image [16, 17]. However, it is unfair to compare the variance of two patches that have significantly different mean intensities. Therefore, the index of dispersion is often used instead, which is the local variation that has been normalized by the mean [16, 17]. It is defined as

$$F_{NVar} = \frac{1}{HW\mu} \sum_x \sum_y (I(x, y) - \mu)^2. \tag{2.2}$$

Furthermore, the histogram of local variances has also been used as a feature. The process begins by applying a log function to compress the histogram, followed by fitting the compressed histogram to a linear model. A smaller slope corresponds to an image with greater intensity variations, suggesting a higher level of sharpness. A pixel in an image is visible if it exhibits a noticeable difference in comparison to its neighboring pixels. Both spatial difference and autocorrelation quantify the degree of similarity between adjacent pixels within a small patch, indicating the level of dispersion of the pixels within that area. Absolute spatial difference [18] can be defined as

$$F_{D\_L1} = \frac{1}{HW} \sum_x \sum_y |I(x,y) - I(x+\delta_x, y+\delta_y)| \tag{2.3}$$

where the offsets $\delta_x, \delta_y \in \{-1, 0, 1\}$ and $|\delta_x| + |\delta_y| > 0$. It is also referred to as the Sum Modulus Difference [19]. By using different combinations of the offsets δ_x and δ_y, we can obtain a total of eight functions representing absolute differences in different directions. However, since noise also contributes to the differences, this feature can only be utilized when it is larger than a specific threshold or on images with minimal noise level. Since the squared distance penalizes large differences more strongly than the smaller ones, using a squared spatial distance could further alleviate the influence of noise [18], which can be expressed as

$$F_{D\_L2} = \frac{1}{HW} \sum_x \sum_y \left(I(x,y) - I(x+\delta_x, y+\delta_y)\right)^2 . \tag{2.4}$$

One of its special cases is the Brenner gradient [20], which is achieved by setting $\delta_x = 2$ and $\delta_y = 0$. Similarly, local "contrast" can also be defined as

$$F_{Contrast} = \frac{1}{HW} \sum_x \sum_y \sum_{\delta_x \in \{-1,0,1\}} \sum_{\delta_y \in \{-1,0,1\}} |I(x,y) - I(x+\delta_x, y+\delta_y)| \tag{2.5}$$

While the fine structures are treated equally for pixels having the same deviation from the mean in the difference measures, the difference of autocorrelation [21, 22, 23] takes the deviation into account. The difference of autocorrelation can be defined as

$$F_{DAC} = \frac{1}{HW} \sum_x \sum_y \left[I(x,y)^2 - I(x,y)I(x+\delta_x, y+\delta_y)\right] . \tag{2.6}$$

The difference between autocorrelation and variance is equivalent if we assume adjacent patches share the same mean and the mean-subtracted adjacent patches are independent. A similar feature is also proposed in [23], which is defined as

$$F_{DAC\_Var} = \frac{1}{HW} \sum_x \sum_y \left[I(x,y)I(x+\delta_x, y+\delta_y) - \mu^2\right] . \tag{2.7}$$

It is claimed that both autocorrelation-based features are robust to noise [23]. Some works make use of the local dynamic range as a feature to measure local contrast [24], which translates to sharpness. The local dynamic range measures the difference between the brightest and darkest pixel within a local patch, which is defined as

$$f_{DR} = \max_{x\in[1,H],y\in[1,W]} I(x,y) - \min_{x\in[1,H],y\in[1,W]} I(x,y). \tag{2.8}$$

To account for the relative dynamic range, a normalized version [24] is defined as

$$f_{NDR} = \frac{\max_{x\in[1,H],y\in[1,W]} I(x,y) - \min_{x\in[1,H],y\in[1,W]} I(x,y)}{\max_{x\in[1,H],y\in[1,W]} I(x,y)}. \tag{2.9}$$

Apart from this, in information theory, Shannon entropy describes the amount of information in a signal which is defined as

$$F_{Entropy} = -\sum_{i=0}^{255} p_i \log_2(p_i) \tag{2.10}$$

where p_i is the frequency of observing a pixel having intensity i. Here we assume the image is 8-bit grayscale. It has been used to detect image sharpness because textured image patches tend to have higher entropy than smooth ones [19, 25]. However, since entropy is in nature a randomness measure, an image containing only sensor noise will have higher entropy than a sharp image full of textures. Following a similar approach, [26] calculate the entropy in the normalized Discrete Cosine Transform (DCT) domain. The assumption behind it is that the normalized DCT of an in-focus image tends to be more uniform than an out-of-focus image, which usually exhibits a mode at low frequencies. To alleviate the impact of i.i.d. noise which affects the DCT spectrum uniformly, [26] makes use of the Bayesian Entropy [27] that weights the higher-valued coefficients more than the lower-valued ones. This feature is formulated as

$$F_{BSE} = 1 - \frac{\sum_{\omega+v\leqslant t} |F_C(\omega,v)|^2}{\left(\sum_{\omega+v\leqslant t} |F_C(\omega,v)|\right)^2}. \tag{2.11}$$

where F_C is the DCT coefficient of image I. High-frequency components are ignored to further alleviate the impact of noise. While the DCT transform loses spatial information, Discrete Wavelet Transform (DWT) represents images in terms of functions that are localized both in the spatial and frequency domain. Wavelet filters such as the Daubechies orthogonal wavelet basis D6 have also been used to extract sharpness-related features, which often stay in the high-frequency related subbands, such as HH, HL and LH [28, 29]. The first order statistics of the wavelet transform coefficients in the level-1 HH, HL and LH subands are obtained as

$$F_{DWT\_Mean} = \frac{1}{HW} \sum_{H} \sum_{W} [|W_{HL}(x,y)| + |W_{LH}(x,y)| + |W_{HH}(x,y)|] \tag{2.12}$$

where W is the wavelet coefficient. The second order statistics is defined as

$$\begin{aligned} F_{DWT\_Var} = \frac{1}{HW} \sum_{H} \sum_{W} [& (|W_{HL}(x,y)| - \mu_{HL})^2 \\ & + (|W_{LH}(x,y)| - \mu_{LH})^2 \\ & + (|W_{HH}(x,y)| - \mu_{HH})^2] \end{aligned} \tag{2.13}$$

where μ is the (absolute) mean value of the corresponding W. While higher-order statistics generally correlate better with blur level than lower-order statistics, they usually amplify noise's impact. The efficacy of applying higher-order statistics for improving performance diminishes as the order increases. The usage of higher-order statistics also leads to an increase in computational cost [28]. The work [30] calculates the level of sharpness as the ratio of the lower-frequency channel wavelet coefficients to the middle-frequency ones. They first decompose the image using three levels of wavelet decomposition, resulting in subbands $\{[W^3_{LL}, W^i_{LH}, W^i_{HL}, W^i_{HH}], i = \{1,2,3\}\}$ where i is the level of decomposition. High-frequency channels $W^1_{LH}, W^1_{HL}, W^1_{HH}$ are ignored since they mainly contain noise and provide little discriminative information. The energy ratio can be formulated as

$$F_{ER} = \frac{\|W^3_{LL}\|_1}{\sum_{i=2}^{3} [\|W^i_{HL}\|_1 + \|W^i_{LH}\|_1 + \|W^i_{HH}\|_1]}. \tag{2.14}$$

Edge-based Features In addition to local statistics-based features, it is widely recognized that the HVS is highly adapted to processing structural information in images, such as edges. Out-of-focus blur greatly reduces the intensity of the edges. Hence, it is intuitive to extract gradient and edge-related features to quantify the level of blurriness. This is usually achieved by high-pass or band-pass filtering. The main difference between these features lies in the filters used to detect and quantify high-frequency components. Commonly used edge detection filters include Sobel, Canny, Prewitt, Scharr, Laplacian, DoG, etc. The filtered results may have opposing signs, consequently, when summing them directly, they tend to cancel each other out. In order to address this issue, L_1 and L_2 are often used to determine the absolute strength of the edges. For example, Tenenbaum Gradient [31], or Tenengrad is defined as

$$F_{TG} = \frac{1}{HW} \sum_x \sum_y S_x(x,y)^2 + S_y(x,y)^2 \tag{2.15}$$

where S_x and S_y are the Sobel filters in the horizontal and vertical directions. The sum of the modified Laplacian [32] is defined as

$$F_{SML} = \frac{1}{HW} \sum_x \sum_y |L_x(x,y)| + |L_y(x,y)| \tag{2.16}$$

where L_x and L_y are the Laplacian filters in the horizontal and vertical directions. A similar feature, the Energy of Laplacian [33], can also be defined as

$$F_{EoL} = \frac{1}{HW} \sum_x \sum_y \left(L_x(x,y) + L_y(x,y)\right)^2. \tag{2.17}$$

The DoG filter can also be used to extract edges, which is defined as

$$F_{DoG} = \frac{1}{HW} \sum_x \sum_y \left[G_{\sigma_1} * I(x,y) - G_{\sigma_2} * I(x,y)\right] \tag{2.18}$$

where G_σ is a Gaussian filter with a standard deviation equals σ. One benefit of the DoG filter is that it eliminates the impact of high-frequency details (if $\sigma > 0$), which usually

includes sensor noise. However, the standard deviations σ_1, σ_2 and their ratio $\frac{\sigma_1}{\sigma_2}$ are hyperparameters that need to be defined according to the image. In [34], they set $\sigma_1 = 0$ and $\sigma_2 = 2$. Some works replace the Gaussian filter with a Box filter [35]. Cumulative Probability of Blur Detection (CPBD) [36] studies the human perception of blur around edges for varying local contrast. The probability $P_{\text{BLUR}}(e)$ of detecting blur around an edge e is calculated based on the width of the edge $w(e)$ and a Just Noticeable Blur (JNB) function. CPBD is formulated as

$$F_{CPBD} = P_{\text{BLUR}}(e) = 1 - \exp\left(-\left|\frac{w(e)}{w_{\text{JNB}}(e)}\right|^{\beta}\right) \tag{2.19}$$

where the JNB is the edge width of JNB, which is a function of local variance around the edge [37]. β is a parameter whose value is obtained by ined by least square fitting. Finally, as the name suggests, the CPBD metric is calculated as the cumulative probability of blur detection. An inherent limitation of the edge-based method is its vulnerability to sensor noise. Regardless of whether the image is in-focus or out-of-focus, the sensor noise will be present. After the filtering, the noise will also remain in the filtered image, which can misguide our detection of edges. Another major drawback is the existing filters may not capture the edges or frequency components that are essential for assessing sharpness. This is reasonable since the filters listed above are designed for general natural images instead of digital pathology images.

Knowledge-based FQA Models with Handcrafted Prediction The local statistics-based and edge-based features mentioned above are not an exhaustive list. Nonetheless, many knowledge-based FQA models utilize these attributes to predict the sharpness or blurriness of images. The pioneering works predict the sharpness based on a set of hand-crafted rules, such as applying thresholds on individual features. Hashimoto at al. [38] and Walkowski et al. [39] used the Haralick contrasts [40] which is determined by multiplying the co-occurrence frequencies in the Gray-Level Co-occurrence Matrix (sometimes also referred to as the Gray-Tone Spatial-Dependence Matrix) by the square distance of pixel intensities. This formula assigns higher weights to the co-occurrence frequencies where the pixel intensity changes rapidly, which usually indicates a strong edge. The entropy of co-occurrence frequency is also used as a feature. Ameisen et al. [41, 42] utilized the

maximum local variations F_{Var} (Eq 2.1). Hashimoto et al. [38] measured the sharpness based on the number and width of edges, where the existence of an edge is determined by setting a threshold on the Tenenbaum gradient F_{GT} (Eq 2.15). Similarly, Zerbe et al. also made use of the Tenenbaum gradient F_{TG} (Eq 2.15).

Jimenez et al. [43] performed a detailed investigation on various general Full-Reference IQA (FR IQA) models, including UQA [44], MS-SSIM [45], [46], VIF [47], as well as No-Reference IQA (NR IQA) models BIQI [48], BRISQUE [49] and NIQE [50]. Some sharpness features such as $F_{Contrast}$ (Eq 2.5), F_{BSE} (Eq 2.11), F_{SML} (Eq 2.16), F_{TG} (Eq 2.15), F_{CPBD} (Eq 2.19) are also investigated. They finally choose F_{CPBD}, $F_{Contrast}$, F_{BSE} and F_{TG} and classify a patch as out-of-focus if at least two features are classified as out-of-focus, using a manually determined threshold for each feature.

Knowledge-based FQA Models with Machine Learning Prediction However, establishing manual thresholds that work in all situations is very difficult. Works have quickly transitioned to using machine learning techniques to either explicitly or implicitly model the prediction rules. Gao et al. [51] proposed the first machine learning-based FQA models by boosting 44 sharpness features using a binary AdaBoost classifier [15]. The AdaBoost classifier combines a set of weak classifiers into a strong one to minimize the overall classification error. The classifier was separately trained on two datasets, both with authentically blurred images. The first one is acquired using the z-stack technique, which we will discuss in detail in Sec 2.1.2. The z-stack is generated by scanning with an offset from $-5\mu m$ to $5\mu m$ with an increment of $0.1\mu m$. A total of 77 2048×2048 slices are collected at different offsets for each of the 37 slides. Images with offset values ranging from $-0.7\mu m$ to $0.7\mu m$ are considered in-focus, while others are considered out-of-focus. The second dataset contains 30 in-focus and 30 out-of-focus expert annotated patches scanned at both $20\times$ and $40\times$ magnification levels. The dimension of the patches is 200×200.

Hashimoto et al. [38] predicted the overall image quality of WSIs using a linear combination of sharpness and noise level measures. The parameters of this linear model are determined through regression on a synthesis dataset. In this dataset, out-of-focus images are generated by convolving a sharp image with Gaussian kernels of various standard deviations. The target quality of the regression task is calculated as the MSE between the synthetic out-of-blur image and its corresponding sharp one.

Lahrmann et al. [35] make use of five sharpness features, including F_{TG} (Eq 2.15), F_{DoG} (Eq 2.18), the number of edges detected by the Sobel filters, etc. Then, a SVM with Gaussian Radial Basis Function (RBF) kernel was trained to classify each image into one of the two categories. The training dataset is composed of 800 manually labeled in-focus and out-of-focus image patches extracted from 63 WSIs scanned at 20×. They found that most features they used were separable without the use of a nonlinear kernel.

Sharpa-Lite [1] [34] used a combination of 16 features, including the SH feature from [38], F_{TG} (Eq 2.15), F_{DoG} (2.18), standard F_{DoG}, F_{EoL} (Eq 2.17) of a Gaussian ($\sigma = 0.5$) blurred image, Haralick contrast and entropy [40]. They trained two classifiers with eight features finally selected, one is based on a decision tree, and the other is SVM. While the SVM model generally performs better, the decision tree gives a series of interpretable classification conditions, which help us better understand each feature. The θ measure, which characterizes the degree of separation of two distributions, shows that the Haralick contrast has the best overall discriminatory ability, followed by F_{EoL}, F_{TG} and Haralick Entropy. They also found that some features are more discriminatory for certain stains. For example, F_{TG} performs better for Hematoxylin and Eosin (H&E), whereas F_{TG} is better suited for Immunohistochemistry (IHC). The training dataset consists of $24,000$ in-focus and $24,000$ manually labeled image patches of the size 200×200, which are extracted from 27 histological WSIs scanned at 20×. Campanella et al.

[52] selected 13 sharpness features including F_{Var} (Eq 2.1), F_{DR} (Eq 2.8), $F_{Entropy}$ (Eq 2.10), F_{D_L2} (Eq 2.4), F_{SML} (Eq 2.16), F_{TG} (Eq 2.15), F_{CPBD} (Eq 2.19), two histogram threshold methods, as well as some general ISA metrics [53, 54, 55, 56]. They generate synthetic out-of-focus images by convolving sharp patches extracted from 207 WSIs with Gaussian kernels of five standard deviations (0.8, 1.2, 1.6, 2 and 2.4). They considered two tasks: multi-class classification and regression. For classification, six classes (five classes for blur and one for sharp) are classified by training two models: a random forest consisting $1,000$ trees and a logistic regression model. For regression, the standard deviations of the Gaussian kernels are considered as the target (0 for sharp images), and a random forest model was trained to accomplish this regression task.

[1] The official code is available at https://bitbucket.org/diapath/sharpa-lite/src/master/.

The PSFs of microscopes in WSI scanners are in nature low-pass filters that attenuate high-frequency information. More recently, FQPath [2] [57] proposed to boost the high-frequency components in visual signals in a balanced way, a well-known functionality found in the HVS. They achieved this by using a symmetric Finite Impulse Response (FIR) kernel [58], which is an approximated inverse PSF of a general WSI scanner. In the first step, a general PSF that corresponds to the out-of-focus optics in the WSI scanners is estimated. Then, the inverse of the estimated PSF is synthesized as a superposition of multiple even-derivative filters. Finally, this inverse PSF was used to extract sharpness-related features. A general purpose ISA model, HSV-MaxPol [3] [2], was developed with a similar motivation. The difference is that the FIR kernel in HSV-MaxPol was designed as the inverse of a Generalized Gaussian kernel which is derived based on NSS.

In spite of the fact that machine learning models assist us in better comprehending and making better use of handcrafted features, they usually result in sub-optimal performances since the feature design is not guaranteed to match our final goal, which is FQA.

Data-driven FQA

As deep learning continues to advance at a rapid pace, CNN is becoming an end-to-end solution that encompasses both the feature design and final prediction processes. This makes CNN a more ideal choice for a variety of image processing applications, including FQA. MIQ [59], DeepFocus [60] and Campanella et al. [52] made the first attempts toward data-driven FQA using CNNs in 2018. DeepFocus [4] [60] proposed a CNN consisting of five convolutional layers, three max pooling layers, and three fully connected layers. The input patch size is 64×64 and the final output is two scalars indicating the probability of the patch being in-focus and out-of-focus, respectively. The network was trained using a cross-entropy loss function, optimized through the Stochastic Gradient Descent (SGD) optimizer. The training dataset [5] is acquired through z-stack, where nine focal plane offsets are used: $-2.5\mu m, -2.0\mu m, -1.5\mu m, -1.0\mu m, -0.5\mu m, 0.0\mu m, 0.5\mu m, 1.0\mu m, 1.5\mu m, 2.0\mu m, 2.5\mu m$. Slices

[2] The official code is available at https://github.com/mahdihosseini/FQPath.
[3] The official code is available at https://github.com/mahdihosseini/MaxPol.
[4] The official code is available at https://github.com/cialab/deepfocus.
[5] The dataset is available at https://doi.org/10.5281/zenodo.1134848.

with an offset in $\{-0.5\mu m, 0.0\mu m, 0.5\mu m\}$ are considered as in-focus while others are considered out-of-focus. A total of 16 WSIs are used to generate the dataset and they choose 2500 non-overlapping patches of 64×64 from each slide. Finally, $120,000$ in-focus patches and $108,000$ patches are used during training and validation. Random horizontal and vertical flipping, as well as random rotation by 90° are used as data augmentation techniques to increase the number of training samples. To evaluate the transferability of the trained model on unseen data, they collect another two datasets for testing. The first testing dataset, consisting of six slides, is generated by the same z-stack method using the same scanner as the training dataset. This testing set consists of $654,912$ 64×64 patches scanned at $72,768$ locations of nine focal distances. The other testing dataset, consisting of two slides, is acquired with a different scanner at a different facility where the patches are manually labeled.

ConvFocus [61] employed a relatively lightweight network, which consists of the first few layers of the Inception V3 network [62], including three convolutional layers, one max pooling layer and one average pooling layer. The training data was generated using a synthetic procedure that simulates the real-world image capturing process in WSI scanners. In the first step, they manually inspected and collected $166,000$ 300×300 sharp patches from $26,526$ slides scanned at $40\times$. In the second step, two types of blur kernels are used to convolve with the sharp images: Gaussian kernel and Heavyside kernel. The 2D Heavyside side kernel is known to approximate the appearance of defocus blur better [63], which is known as bokeh, based on human observations. A total of 29 out-of-focus levels are generated by varying the standard deviation and radius in the Gaussian and Heavyside kernels, respectively. However, the synthetically blurred images looked over-smooth compared to real-world out-of-focus ones. They assume this is caused by the blur operation hiding the high-frequency information, such as sensor noise and the blocking artifact caused by JPEG compression. Therefore, in the third step, they further apply JPEG compression with quality ranging from 70 to 90. In the final step, they add Poisson noise of varying intensities to simulate the sensor noise generated by different WSI scanners. To test the model's ability on authentic out-of-focus images, they collected 7 slides scanned at $40\times$ for testing. A total of $37,715$ patches are extracted and manually labeled with one of the 13 grades, ranging from 0 (in-focus) to 6 (very strong out-of-focus). Besides this,

they further collected a testing set using z-stack. A single slide is scanned at 21 focal planes with a $0.4\mu m$ increment between adjacent planes. All images are cropped to 139×139 before being fed into the network.

Some works also made use of more complex and deeper networks. Campanella et al. [52] assumed that color is independent of FQA for digital pathology. They trained a ResNet18 [64] model from scratch with grayscale images of the size 224×224. Details of the training dataset have been described in Sec 2.1.1. They trained this network separately in two settings: 6-class classification and regression. For classification, cross-entropy was used as the loss function, while MSE was used in the regression setting. Simialar to [52], MIQ [59] also trained a CNN using 84×84 grayscale images. The network consists of two convolutional layers, two max pooling layers, and two fully connected layers. They synthesized the out-of-focus images by convolving the sharp images with PSFs approximated at 10 different out-of-focus focal planes with $2\mu m$ increments. Similar to ConvFocus [61], Poisson noise was also applied to account for image sensor offset and gain. The output of the network is a probability distribution over the 11 defocus levels. What differentiates this model from any other model is the loss function. Most classification tasks use cross entropy as the loss function, where the categories are assumed to be nominal (unordered and mutually exclusive). However, in the case of FQA, categories representing different focus levels are ordered. We refer to this kind of variable as ordinal. To make use of the order information, MIQ [59] used a ranked probability score [65] as the loss function. A measure of classification uncertainty was also computed as the Shannon entropy of the output distribution. Following the same motivation, [66] investigated seven CNNs trained with five loss functions that make use of the ordinal information. These loss functions include ordinal encoding [67], binomial unimodal ordinal encoding [68], a regularized cross entropy and ordinal entropy [69]. The training was conducted on the FocusPath dataset [2], which is acquired based on z-stack, where 16 focal planes are used for capturing images. A detailed description of this dataset will be provided in Sec 2.1.2. It was shown that the ordinal encoding [67] loss function results in the best overall performance across various network architectures and evaluation metrics. The best-performing combination is MobileNetv2 [70] and the ordinal encoding loss function. The major difference between ordinal classification and regression is that the distances between the classes are unknown for

the ordinal case. Some datasets categorize blur levels into five categories: very poor, poor, ok, fairly good, very good. In this scenario, ordinal classification might be handy since these categories are ordered but the distance between classes is hard to define. However, since the distance between any blur levels in a z-stack-based dataset is clearly defined, the advantage of using ordinal classification over regression on z-stack-based datasets needs further investigation.

More recently, Liao et al. [71] studied the FQA problem in the microscope autofocus setting, where a lower magnification at 10× was considered. They selected MobileNetv3 [72] for its optimal balance between performance and computational cost. MobileNetv3 is an enhanced version of MobileNetv2 [70], which was utilized in [66] as described previously. The input size was changed to 672×672 to cover a larger field of view for reliable prediction. The training dataset was collected following the z-stack method, where 25 slices are acquired for each slide, ranging from $-36\mu m$ to $36\mu m$ with a $3\mu m$ increment. The network was trained using a Smooth L1 loss where the target is the defocus distance. To search for the best hyperparameters such as learning rate, momentum, weight decay, etc, population-based training [73] was used.

2.1.2 Focus Quality Assessment Datasets

Labeled data are essential for training either traditional machine learning-based or data-driven FQA models in a supervised learning manner. As far as we know, we did not notice any FQA models designed with an unsupervised learning-based approach. Following the tradition in IQA, labeled datasets can also be categorized into two major types: authentically distorted ones and synthetically distorted ones. For authentically distorted images, the WSIs are captured with real out-of-focus lenses, either intentionally or unintentionally. Other distortions such as sensor noise and compression artifacts are also automatically added to the image in the image processing pipeline within the scanner. Manual labeling and z-stack are the two major methods to create a labeled FQA dataset with authentic distortions.

The manual labeling method follows a more traditional and time-consuming way of collecting both in-focus and out-of-focus patches with their ground truths. A detailed data

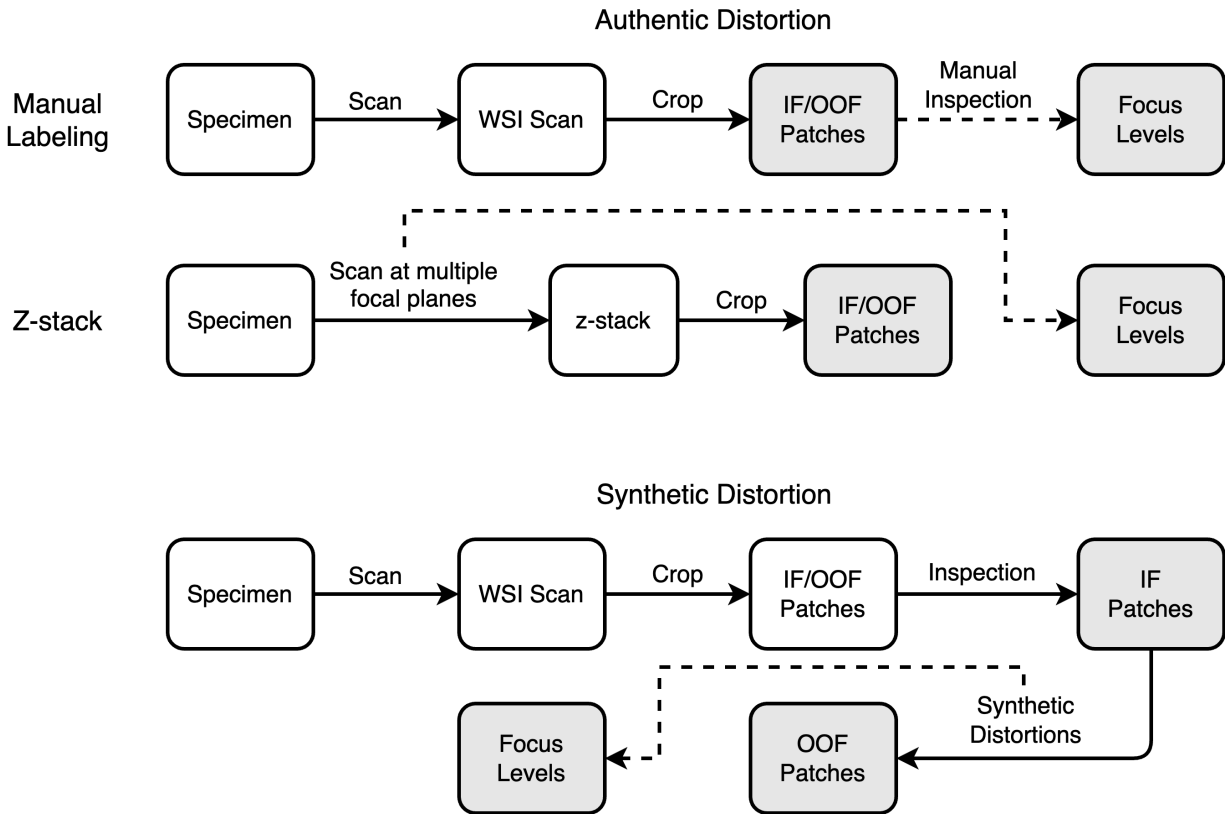


Figure 2.1: The pipelines of major ways of collecting images and focus level labels for FQA. IF is the abbreviation of in-focus and OOF represents out-of-focus. Rounded rectangles filled with gray are the desired output.

collection pipeline is illustrated in the first row of Fig 2.1. In the first step, WSIs are cropped into smaller patches and only the patches that contain tissue are saved for human inspection. This filtering process is mainly achieved through setting a threshold on certain features, such as the mean or variance (Eq 2.1) of pixel intensity. In the second step, trained pathologists are invited to grade each patch into different levels of blur. Due to HVS's limited ability to distinguish the fine-grained level of blur, most works categorize patches into two classes, either in-focus or out-of-focus. Nonetheless, more than two levels of focus are annotated in some works. Due to the tedious procedure of manual annotation, the number of images in this type of dataset is often limited. However, a large amount of

data is required for training CNN-based FQA models.

To bridge this gap, z-stack has become a popular way to generate authentically distorted images with higher efficiency. A detailed data collection pipeline is shown in the second row of Fig 2.1. In the first step, a built-in autofocus algorithm is used to generate a 3-D focus map for the entire specimen as described in Fig 3.1. This focus map is then offset by a series of $N-1$ distances, both toward and away from the objective lens. After this, the slide is scanned according to the original focus map as well as the offset ones. We refer to the N WSIs as the z-stack. An example of the z-stack is provided in Fig 2.4. In the second step, patches are extracted from the stack. The focus level is determined as the offset distance. Since the z-stack is a built-in function in many WSI scanners [74], this method does not require too much labor and is often time efficient.

Even though the z-stack method does not need human annotation, it still requires access to WSI scanners and slides. To make the data acquisition even more efficient, synthetic distortions are used to simulate the appearance of authentic out-of-focus images. This method can make use of the large amount of high-quality unlabeled WSIs that are publicly available. For example, the Cancer Digital Slide Archive (CDSA) [75] hosts more than $30,000$ WSIs sourced from The Cancer Genome Atlas (TCGA), and the number is growing. The synthesis procedure is illustrated in the third row of Fig 2.1. The first step is to extract patches from the WSIs and remove the patches that are out-of-focus or do not contain tissue. Although most of the WSIs in TCGA are of high quality, certain areas within these WSIs may still be out-of-focus due to factors like tissue folding or air bubbles. These blurry areas may not impair the diagnosis, but they will impact the distortion synthesis if not removed. Some works [52, 61] manually checked the sharpness of each patch, while some works [59] select the sharpest patch from a z-stack based on some handcrafted features, such as variance (Eq 2.1). In the second step, blur kernels of various intensities are applied to the sharp patches. Commonly used filters include Gaussian [52, 61, 38], estimated PSF [59] and Heavyside [61]. However, applying these low-pass filters alone may create an over-smooth appearance, which does not match the look of real out-of-focus images. This is partially caused by these low-pass filters removing the sensor noise or compression artifact in the sharp images, which persist in authentic out-of-focus ones. As a result, some works add Gaussian noise [38], Poisson noise [59, 61] and

Dataset	Year	Type	Mag	Scanner	# Slides	# Organs	# Stains
DeepFocus [60] [6]	2018	Z-Stack	40×	Aperio ScanScope	16	N.A.	4
FocusPath [2] [7]	2019	Z-Stack	40×	Huron TissueScope LE1.2	9	9	8

Dataset	# Patches	Patch Size	Pixel Size	# Z-Levels	Focus Range	Focus Step
DeepFocus	$204,000$	64×64	$0.2461\ \mu m$	9	$-2.5\ \mu m \sim 2.5\ \mu m$	$0.50\ \mu m$
FocusPath	$8,640$	1024×1024	$0.2500\ \mu m$	16	N.A.	$0.25\ \mu m$

Table 2.1: The details of the two public available FQA datasets.

JPEG compression to the filtered images to better simulate the signal processing pipeline in the scanners. Finally, the intensities of these blur kernels, often expressed as standard deviation or radius, are used to generate the level of focus.

Although labeled datasets are essential for training either traditional machine learning-based or data-driven FQA models in a supervised learning manner, only two datasets are publicly available, both acquired using the z-stack method. Detailed specifications of these two datasets are provided in Table 2.1. Without confusion, we refer to the training dataset used in the DeepFocus [1] paper as DeepFocus. In DeepFocus, 2500 non-overlapping patches of the size 64×64 are extracted from each of the 16 WSIs. These slides are stained using four different stains: H&E, Ki67, CD21, and CD10. For each patch, another eight patches are captured using the z-stack method by setting an offset at $-2.5\mu m, -2.0\mu m, -1.5\mu m, -1.0\mu m, -0.5\mu m, 0.5\mu m, 1.0\mu m, 1.5\mu m, 2.0\mu m, 2.5\mu m$. This results in a total of $16 \times 2,500 \times 9 = 360,000$ patches. $120,000$ patches with an offset value in $-0.5\mu m$, $0.0\mu m$ and $0.5\mu m$ are considered in-focus since they are visually indistinguishable by experts. $240,000$ patches with other offsets are considered out-of-focus. To balance the number of in-focus and out-of-focus images, $108,000$ out-of-focus patches are sampled from the original set. The publicly available version consists of $204,000$ patches of eleven offset values. The focus level distribution is illustrated in Fig 2.2. Following the rule described

[6] The training dataset is available at https://doi.org/10.5281/zenodo.1134848.

[7] A subset containing 864 patches is available at https://sites.google.com/view/focuspathuoft/. The full version will be made available upon request.

earlier, a total of $10,800$ patches are labeled as in-focus while $96,000$ are out-of-focus. Sample images captured at different focus levels are shown in Fig 2.3. The relatively small patch size 64×64 may limit its applicability in some scenarios. Firstly, some popular CNN architectures, such as ResNet [64], DenseNet [76] and MobileNetV3 [72], require 224×224 input. Secondly, assigning one z-level to a large patch/strip in the autofocus map may not accurately reflect the focus level of smaller patches within it. This is possibly caused by the nonuniform height distribution within the large patch, or the lack of textural details in this small patch. Randomly drawn examples are shown in Fig 2.3, the level of blur of some patches is hard to distinguish by the human eye, even though they are captured at different (absolute) z-levels. On the other hand, the level of blur is more distinguishable for larger patches, as shown in Fig 2.4.

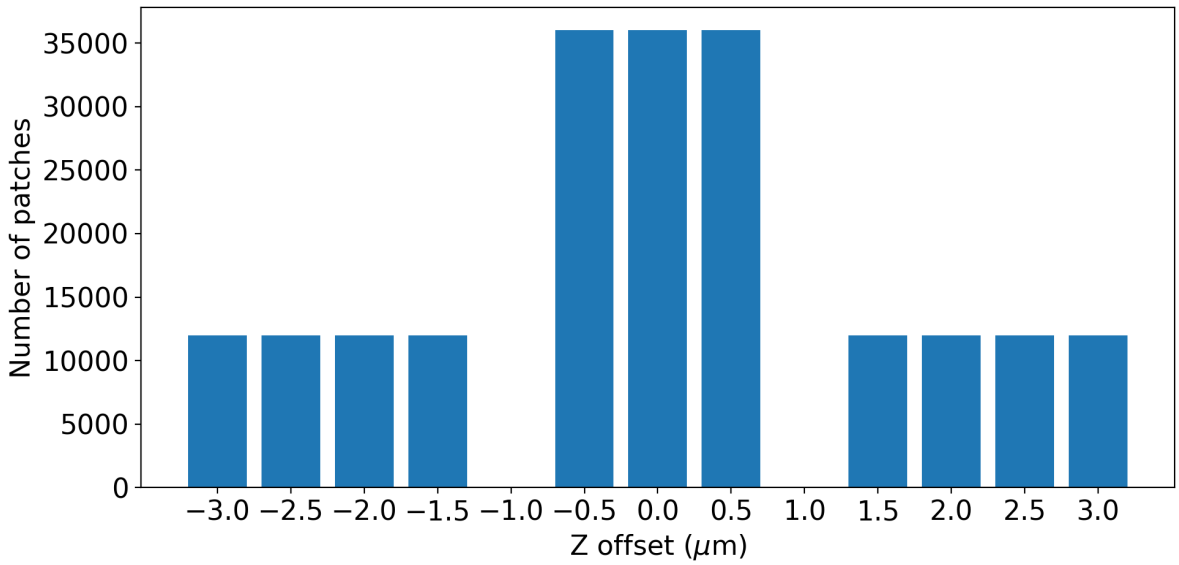


Figure 2.2: Focus level distribution of the DeepFocus dataset [1].

The FocusPath[8] dataset [2] contains $8,640$ patches of 1024×1024 images. To create a database with high diversity, the slides used are cut from nine distinct types of organs with eight types of stains. The stains being used are Trichrome, H&E, Mucicarmine, IRON(FE),

[8]The data is available at https://zenodo.org/record/3926181

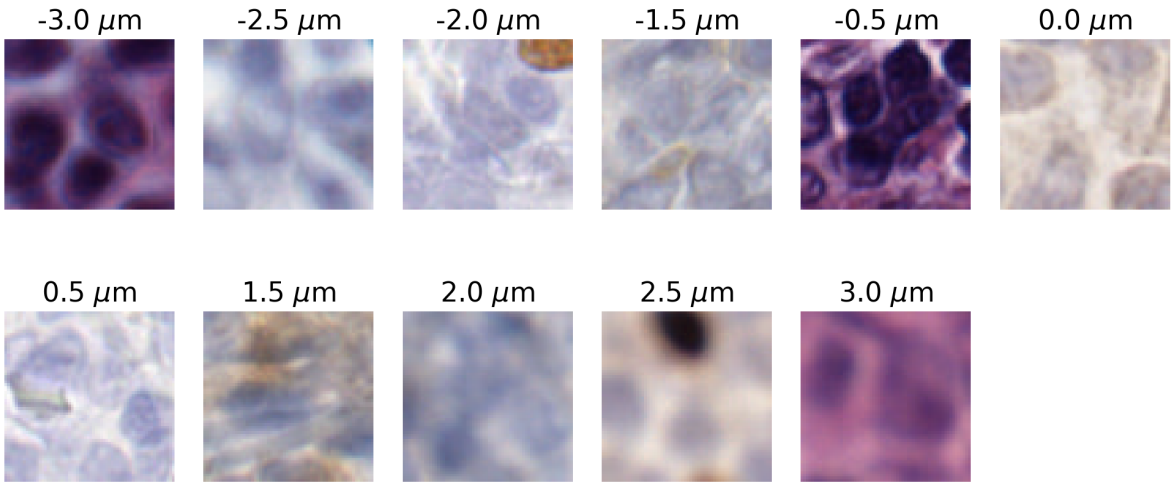


Figure 2.3: Sample images of the DeepFocus dataset [1] at different focus levels.

AFB, Congo Red(CR), PAS and Grocott. In Figure 2.5, randomly drawn examples are shown for each of the slides. The WSIs are scanned by Huron TissueScope LE1.2 [77] using 40X optics lens at 0.25μm/pixel resolution. By using the z-stack method, images are captured at 16 focus levels with an increment of 0.25μm. The z-stack examples are shown in Figure 2.4. Although the number of patches $(8,640)$ in the FocusPath dataset is less than the DeepFocus dataset $(204,000)$, the patch size is substantially larger in FocusPath (1024×1024) than DeepFocus (64×64).

2.1.3 Image Sharpness Assessment Models

Although FQA seems to be a special case of the more general task ISA, digital pathology images differ from general photographic images in several ways. Firstly, most pathology slides are shift and rotation-invariant and can have varying contrast, colors, and textures depending on the staining process, tissue type, and disease state. Secondly, the optical design of the microscopes in WSI scanners differs substantially from general cameras.

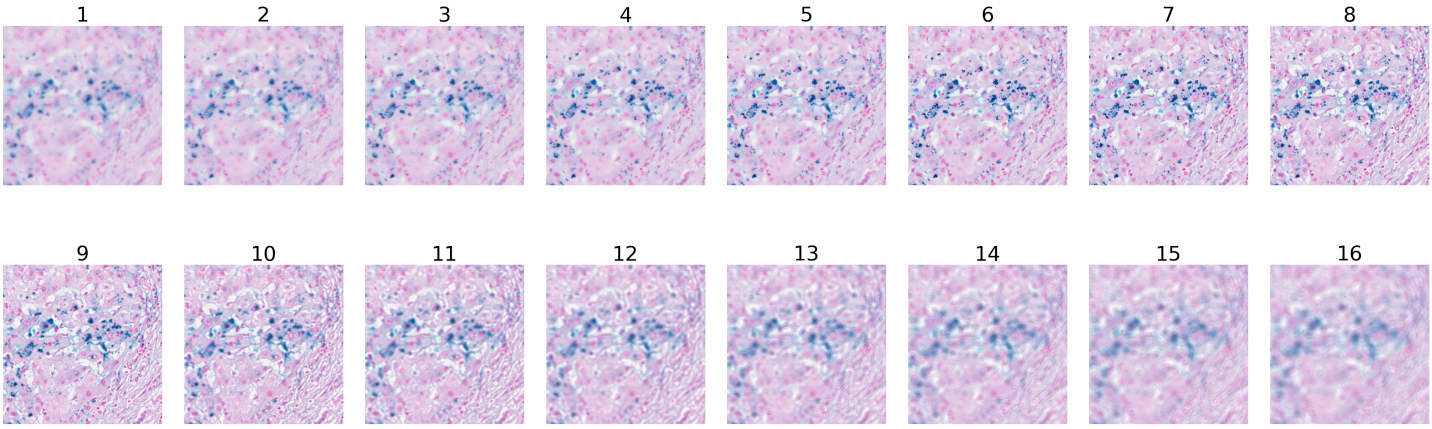


Figure 2.4: A sample of Z-stack images in the FocusPath dataset [2]

.

Microscopic optics are specifically designed to resolve microscopic details at the cellular or even molecular level that are not visible to the naked eye, requiring lenses that can achieve high resolutions with minimal aberrations. This is usually achieved through using a lens with a high magnification rate, high Numerical Aperture (NA) and medium with a higher reflective index than air. For example, the NA for a "large" aperture (F1.8) photography lens is only 0.27. However, common NA for the 40× microscopic lens used in digital pathology ranges from 0.75 to 1.2 [74]. As a result, the two optic systems often have substantially different spherical abbreviation patterns. The most noticeable difference is that the microscopic lens usually has a very shallow DOF, which means that only a very thin plane of the sample is in-focus. Thirdly, the microscopes have controlled lighting conditions, often through transmitted light from below the sample. This controlled lighting system provides constant and even illumination across the entire sample, which is crucial for capturing the details in specimens without interference from outside illumination. On the other hand, consumer cameras rely on ambient light or flashlights which are more variable and less controlled, which may impact the observation of sharpness and contrast. Fourthly, the image sensor and image processing algorithms used are different. WSI scanners use

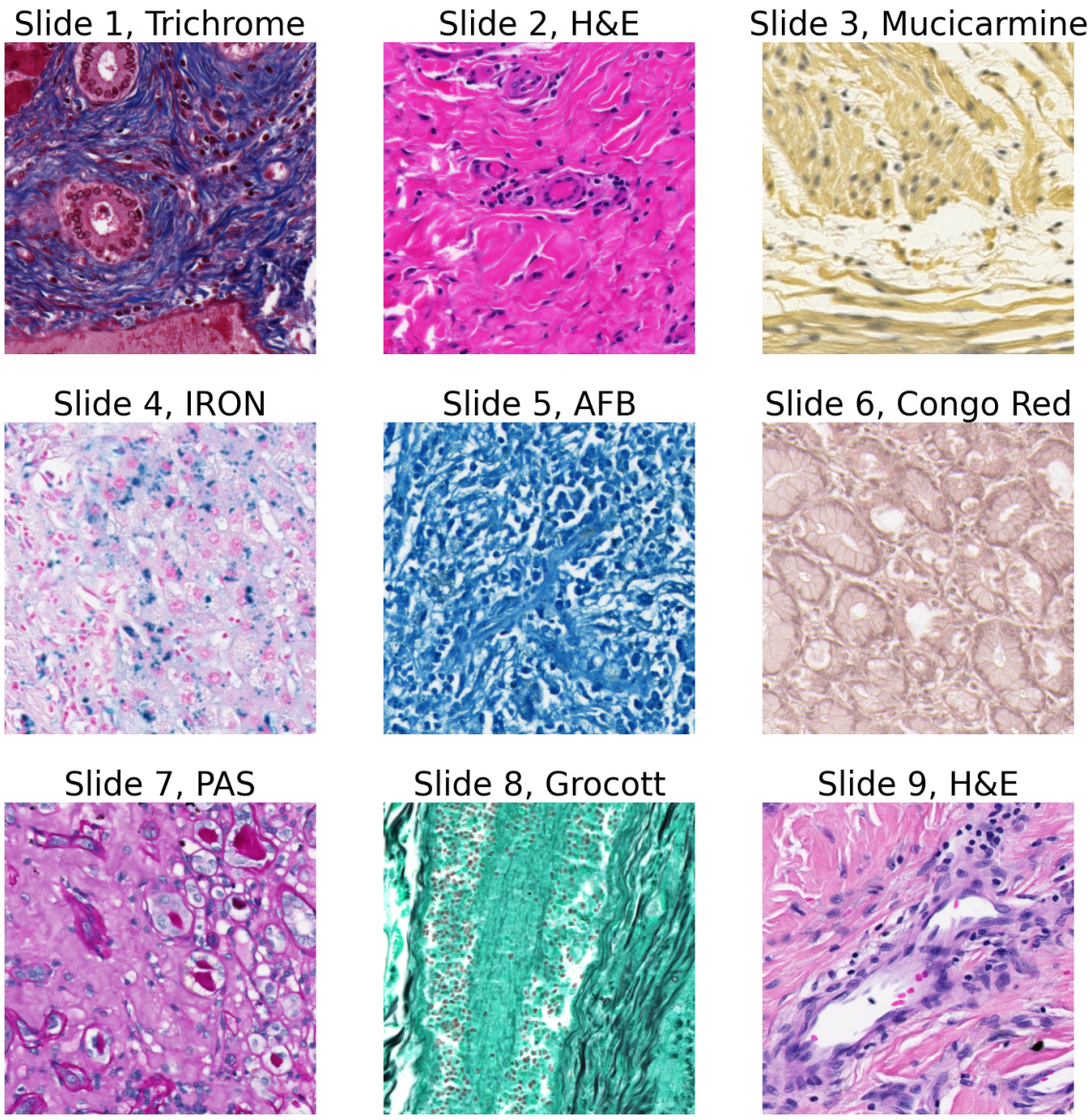


Figure 2.5: Sample images cropped from each of the slides in the FocusPath dataset [2]

.

sensors that are optimized for color accuracy and resolution to account for the fidelity of

stains and detailed structures necessary for diagnostic purposes. The scanned images often have different black-level, pixel gain, photon noise and dynamic range compared to general image sensors [59].

In summary, digital pathology images are distinct from typical photographic images due to differences in the object's physical characteristics, optical systems, illumination conditions, imaging sensors, and the image processing techniques employed for capture. General ISA models are designed to work with general photographic images that may not align well with the specific needs of digital pathology images. Nevertheless, this section presents a thorough literature review of ISA models to highlight both the similarities and differences between FQA and ISA, thereby enhancing the comprehension of their applicability and limitations in the context of digital pathology.

Unlike in FQA, where only no reference models are available, ISA models can be either no reference or full reference. In order to stay in line with our topic, our review will only consider no reference ISA models in this section. In the following, we will be referring to no reference ISA as ISA for simplicity. Similar to the classification method utilized in the review of FQA, ISA can be divided into two classes based on the source of the feature, namely knowledge-based and data-driven-based. Knowledge-based methods can be further categorized into edge-based, local statistics-based and NSS-based.

Knowledge-based method When an image is blurred, most low-frequency components remain while high-frequency components are suppressed. Consequently, the appearance change of edges in blurred images will be more significant than in smooth regions. In other words, edges are more prone to blur. By examining some properties of the detected edges or gradient, we can estimate the sharpness of the image. The first step is edge detection. Commonly used edge detection filters include Sobel, Canny, Prewitt, Scharr, Laplacian, DoG, etc. The most commonly used edge-based feature is edge width [78]. By taking the properties of the HVS into account, some works considered the impact of local contrast [37, 36] and contrast of the edge [79] on the visibility of edge width. The entropy of edge [80] map was also used to estimate the sharpness. Besides deducing the feature to a sharpness score based on a handcrafted formula, [81] regresses the edge features to subjective evaluated scores using a power function. [82] predict the image sharpness using a weighted geometric mean of the maximum gradient and the variability of gradients.

[83] measured sharpness using perceptually weighted statistics of image gradients. Some methods explicitly model the blurred edge by convolving a parametric edge model with a blur kernel. Most models assumed the edge to be an ideal step function and the blur filter to be a Gaussian kernel. By estimating the parameters of the kernel, we can assess the blurriness of the edge [84, 85, 86, 87]. Similarly, Wu et al. [88] assume the blur kernel to be a line spread function (LSF), which is essentially a 1-d version of PSF. They estimate the parameters of the LSF from a blurred edge. Then the 2-d PSF, which is represented by a uniform disk with radius R, is inferred from the LSF. Liu et al. [89] proposed a local statistics model of edge width based on two observations: 1) the histogram of edge width shifts to the right and 2) tends to spread out when the level of blur increases. The major drawback of edge-based methods is that edge detection tends to be inaccurate with the presence of blur. Besides using handcrafted formulas to summarize the edge-related features into sharpness scores, some models [90, 91, 92] use machine learning techniques, such as SVR, for prediction.

Besides extracting edge information, some models measure sharpness based on local statistics. For local statistics in the spatial domain, MLV [93] measured the maximum local variation of each pixel with respect to its eight neighboring pixels. This feature is defined as

$$F_{MLV} = \max_{\delta_x, \delta_y} |I(x, y) - I(x + \delta_x, y + \delta_y)| \tag{2.20}$$

where $\delta_x \in \{-1, 0, 1\}$ and $\delta_y \in \{-1, 0, 1\}$. F_{MLV} is similar to F_{D_L1} (Eq 2.3), which is one of the features used in the hybrid ISA model S_3 [94]. Instead of computing the average as in F_{D_L1}, F_{MLV} finds the maximum value, which captures the variations better than F_{D_L1}. Finally, the standard deviation of the F_{MLV} distribution was used as a feature to measure sharpness. Gu et al. proposed the ARISM model [95], which is inspired by the free energy principle [96]. The foundational assumption is that the brain tries to explain the scene using an internal generative model, which is molded using an auto-regressive (AR) model here. According to the observation, image blur increases the similarity of locally estimated parameters of the AR model. ARISM measures the energy-difference and contrast-difference of the coefficients of the AR model at each pixel individually and com-

putes the image sharpness utilizing percentile pooling. BIBILE [97] calculate the discrete Tchebichef moment of edges, which is effective in describing the shape of the whole image. The sharpness score is determined based on the variance-normalized moment energy of the moments, weighted by a saliency model to align with human visual system characteristics. In [98], Deng et al. modeled the distribution of the gradient magnitude as Weibull [99]. Divisive normalization [100] is also utilized to reduce the influence of image content and achieve more robust performance. The final sharpness score is predicted using a Sparse Extreme Learning Machine that uses both L_1 and L_2 regularizations [101, 102].

Both edge-based and gradient-based methods are usually susceptible to noise. By calculating the local statistics in the frequency domain, we can partially separate the edge from the noise. For example, the Lipschitz regularity theory is widely used to separate the edge and non-edge (usually noise) coefficients of the Dyadic Discrete Wavelet Transform (DDWT). Based on this theory, Ferzli et al. [103] applied a 3-level DDWT to a (noisy) image, the edges can be separated from the noise. By calculating the edge width [78] on these edge maps, one can estimate the sharpness of an image that may contain noise. Vu et al. proposed a fast image sharpness index FISH [104]. The image was first decomposed using three level of discrete wavelet transform. The sharpness is calculated as a weighted average of the log energies of the wavelet coefficients. [105] further extend the FISH model using saliency maps to account for the nonuniform distribution of blur across an image. Wang et al. [106] discovered that edges induce strong local phase coherence structures in both scale and space within the wavelet domain, while blur usually leads to a deterioration of this phase coherence. Based on this observation, Hassen et al. further developed the sharpness measure LPC [107]. Similarly, Global Phase Coherence (GPC) [108, 109] quantifies the impact of phase information destruction on the total variation of an image. It is observed that sharp images exhibit a higher sensitivity to phase distortions than blurred images, thereby establishing a clear relationship between GPC and image sharpness. Caviedes et al. [110] examined the normalized power spectrum of DCT coefficients near image edges and assessed sharpness through kurtosis calculations. Blurred images are known to exhibit shorter power spectrum tails due to the attenuation of high-frequency components, making power spectrum tailedness a powerful sharpness metric. Considering the fact that the viewing distance affects the perceived sharpness, Li et al. proposed a multiscale model

RISE [111] that utilized features both in the spatial and frequency domain, including multiscale DCT entropy, gradient similarly and singular value similarity. The sharpness score is deduced from these features by training a SVR model. Sang et al. [112] proposed a sharpness index by measuring the falloff pattern of the singular value curve. Similar to the power spectrum plot, the singular value curve becomes increasingly steep (shorter tail) with more blurriness. The sharpness is estimated by fitting the curve to an inverse power function. [58] presented a sharpness metric that utilizes the MaxPol convolution kernels to approximate the first and third-order image differentials. The kernels are regulated at higher cutoff frequencies to balance information loss and noise sensitivity. Adaptive thresholding was applied to the convolved image and the m^{th} central moment was taken as the sharpness score. Based on the observation that the HVS boost high-frequency components, [2] proposed a way to estimate sharpness by simulating this functionality of the HVS. They modeled this behavior as the inverse of a Generalized Gaussian kernel, and it was approximated using the MaxPol kernels. Some methods [94, 111, 113] have also combined features from both spatial and frequency domains for sharpness estimation. [114] use a MLP with one hidden layer to classify the sharpness of an image into five categories. A total of eight features in both the spatial and frequency domains, as well as some sharpness metrics are used to feed the input neurons. Besides out-of-focus blur, [115] studied motion blur introduced by shaking camera using spectral analysis.

Different from local statistics-based methods that assess sharpness based on features within one image, NSS-based methods rely on the inherent statistical properties observed in natural scenes. NSS models assess the naturalness of a distorted image by quantifying its deviation from the statistical patterns. SPARISH [116] constructed an overcomplete dictionary using natural images, where most entries are edge patterns that are similar to the simple cells in the primary visual cortex [117]. The blurred image is first represented using sparse coefficients in a block manner. The sharpness score is defined as the variance-normalized energy over a set of selected high-variance blocks. Similarly, Lu et al. [118] introduce a multiscale max-pooling on the sparse representation. A SVR is used to map the feature into a sharpness score.

Data-driven method While the aforementioned knowledge-based methods rely on handcrafted features, their effectiveness in capturing the most important features for sharp-

ness assessment is limited. These models mainly target synthetic Gaussian blur, casting uncertainty on their applicability to other blur varieties and authentically blurred images. Designing handcrafted features specific to each blur type is a tedious process. In contrast, with the advancements in deep learning, CNN seamlessly integrates feature design and regression within an end-to-end framework, ensuring optimal results. One major limitation of CNN-based method is it requires a large amount of labeled training data. In the field of ISA or IQA, subjectively rated images are very expensive to collect. Many models come up with ways to mitigate the data shortage issue through unsupervised learning. Zhang et al. [119] pretrained a Siamese network by learning quality ranks among the synthetically blurred images without any subjective ratings. The model can be pretrained on a very large dataset which provides an effective prior for sharpness assessment. The model is then finetuned on small-scale datasets with human labels for evaluation. Based on the observation that the HVS's perception of distortion depends on the image content, Li et al. [120, 121] proposed a NR IQA method that focuses on blur distortion based on semantic feature aggregation to alleviate the impact of image content variation. The semantic feature is extracted using a pretrained image classification model. As most works only addressed synthetic blurs, RBA [122] tried to assess the sharpness of realistically blurred images. Realistic blurs, which can be caused by moving objects, lens aberration, atmospheric turbulence, or camera shaking, are more complex and difficult to characterize. Based on the assumption that the HVS estimate image quality by evaluating the discrepancy between the blurred image and a hallucinated sharp one, RBA developed a feature extraction CNN to predict the discrepancy map using the distorted image. The discrepancy map is generated using sharp images and two types of synthetic blur: defocus blur and motion blur. After the training is done and the features are gathered, the entropy of primitive and variation of power of the features are combined to predict the sharpness.

Autofocus Methods In microscopy imaging, autofocus methods aim to find the best focus plane for the specimen that maximizes the sharpness of the image. To choose the best focus plane, the autofocus models usually capture several images along the optical axis, compare the sharpness, and estimate the optimal focus position. FQA can either work as part of the autofocus pipeline, or in post-capture quality assurance. In the literature, autofocus methods can be categorized into three types: 1) hardware-aided autofocus, 2) z-

stack image-based autofocus, and 3) single image-based autofocus. A lot of hybrid methods are also developed to utilize the strength of each type of method.

The hardware-aided autofocus methods rely on additional hardware or specially designed sensors. Instead of measuring the sharpness of 2D images as in FQA, these methods interact with the 3D specimen and measure the defocus distance of the sample. This type of method includes independent dual sensor scanning, beam splitter array, titled sensor, phase-detection, dual-LED illumination [123]. Due to the additional hardware requirement, WSI systems equipped with this type of autofocusing module are often more complicated and expensive.

The z-stack image-based and single image-based autofocus methods are based on FQA. Generating a 3-D focus map before the scanning is the most common way of autofocus in WSI scanners [124]. The z-stack method is usually adopted to capture multiple images along the optical axis for each point on the map. Then the optimal focus plane is determined by comparing these images. Intuitively, FQA models can be used in this step to measure the sharpness. However, since the FQA model needs to assess the sharpness of each tile at each z-level, determining the optimal focus plane is very time-consuming. This needs the FQA models used to be lightweight and super-efficient. Single image-based autofocus methods often predict the defocus distance based on deep learning. This type of method does not require additional hardware. However, their accuracy and generalizability depend on the quality and diversity of the training dataset.

2.2 Image Quality Assessment Score Fusion

2.2.1 Taxonomy of IQA Models

Image quality assessment (IQA) models aim to predict the perceived image quality by human observers. It has wide applications in the fields of image processing and computer vision. For example, they are used as the evaluation criteria to compare algorithms. In addition, they also serve as guides to drive the design and optimization of perceptually inspired algorithms and systems. Assessing image quality seems to be an easy task for

humans, however, the underlying mechanisms are not well understood, making IQA a challenging task. Image quality is a complicated measurement involving both the image content, context and the HVS. Common attributes contributing to image quality may include image resolution, sharpness, noise level, compression artifact, composition, object recognizability, etc. However, enumerating all attributes is almost impossible, and modeling the interplay between image attributes and the HVS is even more difficult. To evaluate the quality of a distorted image, we need to consider whether or not we have the counterpart pristine image termed as the reference image of the distorted image. Based on the availability of pristine reference images, we can categorize IQA models into three categories: FR IQA, Reduced-Reference IQA (RR IQA) and NR IQA. To evaluate the quality of a distorted image, FR IQA methods require its pristine reference. RR IQA methods require access to certain pre-extracted features of the pristine reference image. On the other hand, NR IQA methods evaluate the quality of the distorted image without any information about the reference image. Another type of IQA model is Degraded-Reference IQA (DR IQA), in which the reference image is a distorted version of the pristine image. In this section, we mainly focus on FR IQA and NR IQA, which are the two most extreme and representative cases in the family of IQA models. The above model taxonomy is created based on the reference image availability, which does not provide much information about the design philosophies of these models. To provide a better understanding of the different principles behind these models, we created a more detailed taxonomy for IQA models, which is illustrated in Fig 2.6. However, due to the distinct design philosophies described above, some models often capture some particular types of distortions or handle some specific image contents better than others. Consequently, individual IQA models often fail to address all types of images and distortions encountered in real-world scenarios.

An intuitive idea is to harness the strengths and mitigate the weaknesses of each IQA model, by fusing the scores of multiple models into a stronger one. In the literature, IQA score fusion methods can be categorized non-mutually exclusively into empirical fusion, rank fusion, and supervised learning-based [125]. Empirical models [126, 127, 128, 129, 130, 131, 132, 133] fuse a predetermined set of IQA models using a handcrafted formula. This approach significantly constrains its adaptability when introduced with new IQA models. Rank fusion methods operate in the discrete rank domain, where the range of

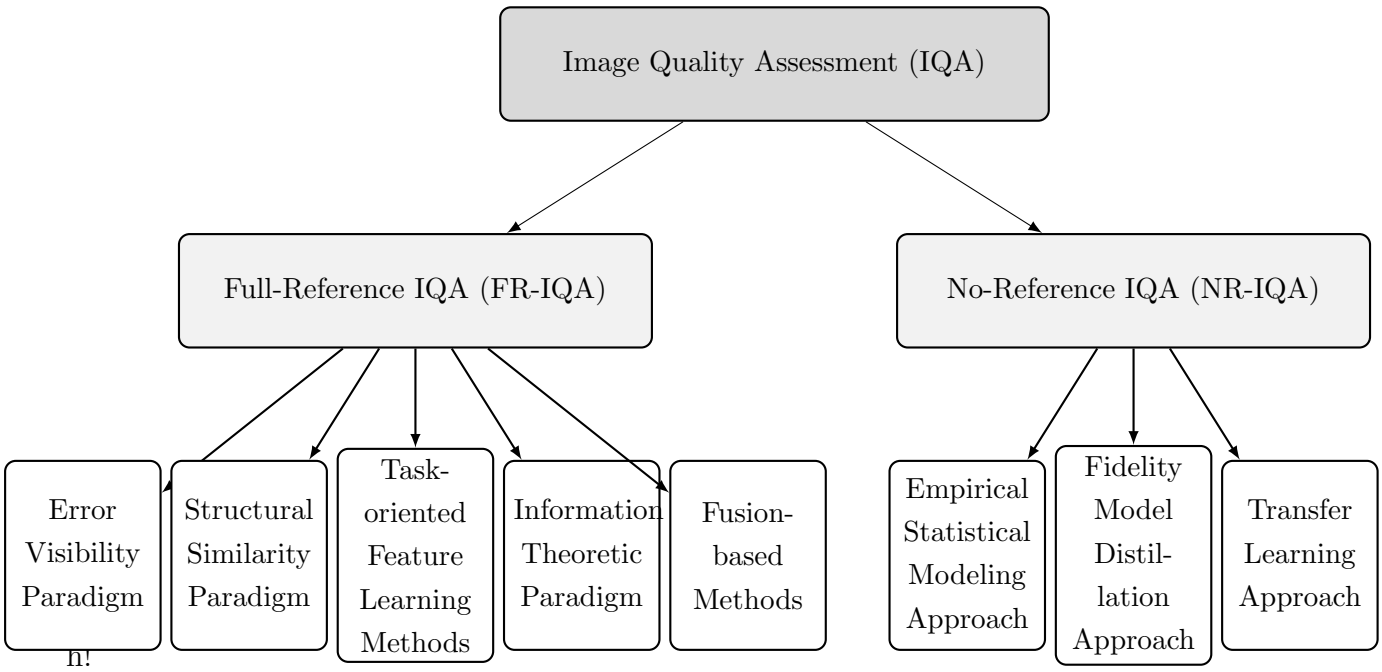


Figure 2.6: A taxonomy for IQA models based on their design principles.

all IQA models is mapped to the same uniform distribution. However, these methods are tied closely to the diversity of the ranking dataset, which can impede generalizability. Supervised learning-based methods [134, 135] are trained under the guidance of the MOS of a single subjective rated dataset. Such fusion methods are essentially refined versions of supervised learning-based IQA models since they share the same ground truth, i.e., MOS of a specific dataset, as the base IQA models. These fusion methods are more closely related to supervised learning-based IQA models since they share the same ground truth, MOS. Nevertheless, these black-box models often suffer from limited generalizability and lack of explainability. In the following, we review each of these IQA fusion methods in detail, as well as some general score/rank fusion methods that are applicable to this task.

2.2.2 Empirical IQA Score Fusion

The first category is empirical score fusion methods. They include earlier methods in IQA score fusion, including HFSIMc [126], CISI [127], CM [128, 129, 130], CQM [131], EHIS [132] and BMMF [133], etc. Utilizing a predetermined set of IQA methods, scores are fused through a handcrafted weighted sum or weighted product formula. The weights are determined either via prior knowledge or optimization on subjectively rated datasets. For example, HFSIMc [126] is defined as a weighted product of RFSIM [136] and FSIMc [5]:

$$\text{HFSIMc} = (\text{RFSIM})^{0.4} \cdot (\text{FSIMc})^{3.5} \tag{2.21}$$

where the exponential parameters are determined by fitting the MOS on the TID2008 dataset [137]. Similarly, CISI [127] is defined as a weighted product of MSSSIM [45] and VIF [47] and FSIMc [5]:

$$\text{CISI} = (\text{MSSSIM})^{0.5} \cdot (\text{VIF})^{0.5} \cdot (\text{FSIMc})^{5}. \tag{2.22}$$

CM3 [128, 129] is defined as a weighted product of IFC [46], NQM [138] and VSNR [139]:

$$\text{CM3} = (\text{IFC})^{0.34} \cdot (\text{NQM})^{2.4} \cdot (\text{VSNR})^{-0.3}. \tag{2.23}$$

CM4 [128, 129] is similarly defined as

$$\text{CM4} = (\text{IFC})^{0.2} \cdot (\text{NQM})^{2.9} \cdot (\text{VSNR})^{-0.54} \cdot (\text{VIF})^{0.5}. \tag{2.24}$$

In [130], the authors conducted a comprehensive grid search of different combinations of two and three IQA models using the weighted product formula. CQM [131] is defined as a weighted product of MS-SSIM [45], VIF [47] and R-SVD [140]:

$$\text{CQM} = (\text{MS-SSIM})^{7} \cdot (\text{VIF})^{0.3} \cdot (\text{R-SVD})^{-0.15}. \tag{2.25}$$

EHIS [132] is defined as a weighted product of MS-SSIM [45], VIF [47], weighted FSIM (WFSIM) [5] and RFSIM [136]:

$$\text{CQM} = (\text{MS-SSIM})^{1.6131} \cdot (\text{VIF})^{0.2037} \cdot (\text{WFSIM})^{59.7151} \cdot (\text{RFSIM})^{0.1989}. \tag{2.26}$$

Different from the above fusion methods that treat all image content and distortion equally, BMMF [133] adaptively adjusts to local image content and distortion. More specifically, the texture of each image patch content is classified into three types: smooth, edge and texture. Based on this information, each patch is further grouped into three classes: simple, normal and complex. To evaluate the distortion of each image, they first extract five statistics: blockiness, average absolute difference between in-patch image samples, zero-crossing rate, average edge spread, and average patch variance. Based on these statistics, each distorted image is grouped into one of the five distortion groups, which contains a total of 17 distinct distortion types. Finally, an image quality metric is selected according to a pre-defined lookup table for each combination of content and distortion. All patch scores are fused to one overall score using a weighted summation, where the weights are determined by fitting the MOS on a small dataset.

Although empirical fusion methods are simple to design, they have several limitations. 1) The handcrafted functional, often implemented as a weighted sum or product, may not be expressive enough to capture the complex underlying patterns. 2) Due to the fact that the exponential parameters are obtained by fitting the MOS in most models, they

also belong to the supervised learning-based models. As a result, they suffer from the same limitations as those supervised learning-based score fusion methods, which we will discuss shortly. 3) They fail to account for the score-dependent uncertainties. Although BMMF [133] takes local image smoothness and distortion features into account, the model selection process and the fusion formula are all handcrafted. This means that the content-dependent uncertainty estimation is pre-determined and fixed. Thus, there is a need for more sophisticated and flexible fusion methods that can address these limitations.

2.2.3 Supervised learning-Based Score Fusion

Instead of using handcrafted fusion formulas, supervised learning-based fusion methods consider the fusion process as a black box. The mapping is automatically learned using machine learning techniques. Models within this category are trained under the supervision of the MOS of a single training dataset. MMF [141, 134] and [135] use SVR while CNNM [142] uses a neural network to learn the mapping from scores to MOS. More specifically, MMF [141, 134] first classifies the distortion type using a similar approach as BMMF [133]. For each distortion type, it fuses 10 IQA models using a SVR with radial basis kernel. Chetouani et al. [135] fuses 4 IQA metrics using a SVR with Gaussian kernel. However, the computational complexity of training a SVR with a non-linear kernel is $\mathcal{O}(n^2 d)$ where n is the number of samples and d is the feature dimension. This poses a big challenge for training a SVR on a large dataset. Chetouani et al. [135] have also experimented with MLP and found out it is inferior to SVR with respect to performance. This might be due to the fact that MLPs have more parameters and need more data to train. The same phenomenon is also observed by Lukin et al. [142], who propose the use of a Cascade-forward network to increase the performance and reduce the computational complexity. The Cascade-forward network is similar to MLP, with the difference that it connects the input and every previous layer to the following layers. One may consider it as a MLP with dense skip connections between non-adjacent layers.

These methods are more flexible than the empirical fusion methods since the mapping function is learned from data rather than handcrafted. However, the major drawbacks of these supervised learning-based methods also come from the training data. The distri-

bution and interpretation of MOS differ substantially across subjectively rated datasets, hampering the generalizability of these models when encountering unseen data. Such a misalignment is mainly caused by the discrepancies in the methodologies and protocols used in these subjective experiments. For example, differences in image selection, participant recruitment and training, rating protocol, data processing, etc. can significantly influence the distribution and physical meaning of MOS. Consequently, it is challenging to align MOS in different datasets. Another issue caused by the misalignment is that it is difficult to train the models on a large combined dataset consisting of several smaller ones. This is a desired property since subjectively rated datasets are generally small in size.

2.2.4 Rank Fusion

Instead of fusing continuous scores, rank fusion methods first convert scores of individual IQA methods to discrete ranks and then fuse them to attain unified rankings or scores. In score fusion, we need to account for the different ranges of these IQA methods, which intensifies the complexity of this task. A notable advantage of converting scores to rankings is that it nonlinearly transforms arbitrary score distributions into the same uniform distribution. In the IQA literature, there is one model that adopts this approach. RAS is an empirical IQA rank fusion method that is introduced in the BLISS model [143] for generating synthetic MOS. RAS fuses the ranks using a handcrafted Reciprocal Rank Fusion (RRF) [144] formula. The RRF score of image I_i is formulated as

$$\mathrm{RRF}(I_i) = \sum_{j=1}^{M} \frac{1}{k + r_i^j} \tag{2.27}$$

where $r_i^j \in \mathbb{N}$ is the rank of the score of image I_i given by the j-th IQA model. k is a constant to stabilize the calculation. Since the ranks need to be evaluated on a dataset, the calculation of a RRF score of a single image still requires calculating all images within a large dataset.

Besides IQA rank fusion, there are some general purposes rank fusion methods that are related to our task. Akritidis et al. propose two weighted rank fusion frameworks [145, 146]

where the weights are used to estimate the model-level uncertainty. The rank fusion method used can be arbitrary, for example, RRF. They iteratively update the weights using a formula until the changes in the weights are less than a threshold. Some methods [147, 148, 149] rely on the Maximum Likelihood Estimation (MLE) framework to unsupervised aggregate the ranks. Most of them rely on a distance measure between two lists of discrete ranks. There are two types of popular distance measures, namely Spearman's distance and Kendall's distance. The Spearman's footrule distance between two permutations of a list consisting of N items is given by

$$\mathrm{SF}(r^m, r^n) = \sum_i^N |r_i^m - r_i^n| \tag{2.28}$$

where r_i^m and r_i^n are the rankings of the i-th item in the m-th and n-th permutations, respectively. Similarly, the Spearman's distance is defined as

$$\mathrm{SD}(r^m, r^n) = \sum_i^N (r_i^m - r_i^n)^2. \tag{2.29}$$

Both distances have a range $[0, \infty)$, which is not normalized. A normalized version is the SRCC, which is defined as

$$\mathrm{SRCC}(r^m, r^n) = 1 - \frac{6 \cdot \mathrm{SF}(r^m, r^n)}{N(N^2 - 1)} \tag{2.30}$$

whose values range from -1 to 1. The Kendall's distance is defined as

$$\mathrm{KD}(r^m, r^n) = \left| \left\{ (i, j) | r_i^m < r_j^m \wedge r_i^n > r_j^n \right\} \right|. \tag{2.31}$$

It counts the pairwise disagreements between two rankings, which is often referred to as the bubble sort distance. Kendall's distance is also not normalized. A normalized version is the Kendall Rank Correlation Coefficient (KRCC), which is defined as

$$\mathrm{KRCC}(r^m, r^n) = 1 - \frac{4 \cdot \mathrm{KD}(r^m, r^n)}{N(N - 1)} \tag{2.32}$$

, which has a range $[-1, 1]$.

Once we have these distance measures, the likelihood $p(r^m|r^n)$ can be molded using the Mallows model [150], which is defined as

$$p(r^m|r^n, \theta) = \frac{1}{Z(r^n, \theta)} e^{\theta d(r^m, r^n)} \tag{2.33}$$

where $d(\cdot, \cdot)$ is a distance measure of chosen, $\theta \leq 0$ is a disoersion parameter, and $Z(r^n, \theta) = \int_r e^{\theta d(r, r^n)}$ is a normalization constant. The distribution is similar to a generalized Gaussian distribution where r^n is the location and θ is the scale. The distribution becomes more concentrated at r^n as θ decreases. In terms of IQA rank fusion, θ can also be considered as a model-level uncertainty parameter: a large θ indicates more uncertainty in r^m, i.e., model m, with respect to its mode r^n. However, this uncertainty estimation only remains coarse-grained at the model level.

Although the likelihood function makes it possible to conduct the rank fusion using the MSE, optimization in the discrete space is time-consuming compared to those in the continuous space. Furthermore, the conversion from score to rank inherently results in information loss, setting an implicit upper bound on the optimal performance achievable by these rank fusion methods. An additional limitation is the accuracy of rankings relies heavily on the diversity of the dataset. If the number of images is small or the qualities of the images are similar, the rankings will not be meaningful. As a result, the fusion results will also be less accurate.

2.2.5 Pairwise Ranking Guided NR-IQA Methods

This class [151, 152] of methods is not directly related to our topic. They predict image quality based on the distorted image itself rather than IQA scores. IQA scores are only used in the training stage to guide the model through pairwise comparisons. The underlying assumption is that the order of the qualities of two images is more robust than the absolute difference between scores.

2.3 Whole Slide Image Defocus Restoration and Synthesis

2.3.1 Whole Slide Image Deblur

Traditional WSI deblurring methods [153] iteratively estimate the blur kernel and restore the image. More recently, deep learning-based WSI deblur models [154, 155, 156, 157] restore the in-focus image in an end-to-end manner. This is achieved by training on paired out-of-focus and in-focus image pairs. These methods mainly differ in the source of the out-of-focus images (captured or synthetic) and the network design. End-to-end Image deblur can be applied in the scenario where the tissue thickness is uniform and the defocus level is homogeneous across the image. In this case, the training image pairs can be easily collected. However, when the defocus level is inhomogeneous, the in-focus regions are distributed across more than one z-level. Consequently, we normally cannot capture the ground truth image that every pixel is in-focus using normal bright field microscopes. This is different from general-purpose natural image deblur where the in-focus target can be captured using a smaller aperture or by the DOF fusion technique.

The traditional whole slide image deblur model is based on conducting blur kernel estimation and deconvolution either sequentially or iteratively. [153] presentd LB-DVD for restoring out-of-focus fluorescence microscopy images. The method utilized inhomogeneous deconvolution to adaptively restore the non-uniformly defocused image. This is achieved by restoration at the patch level. In the first step, the model estimates the defocus map of the image using an out-of-focus level estimation network called DelpNet. DelpNet is trained on a synthetic dataset and during inference, it classifies each 84×84 patch into one of the twelve defocus levels. Once the defocus level is determined, it deconvolves the image using a parametric PSF, which is a function of the defocus level, following the Richardson-Lucy deconvolution method [158]. The major drawbacks of this model are three-fold: 1) DelpNet is trained using synthetically generated out-of-focus images with an oversimplified distortion model. Consequently, it might not capture real-world out-of-focus patterns of various tissues and microscopes. 2) The deconvolution process uses a parametric PSF model and assumes the noise distribution to be Poisson, which might not match the one

in real-world applications. 3) Due to the non-invertible PSF and the presence of noise, the deconvolution process can generate artifacts.

The deep learning-based whole slide image deblurring models addressed the challenges in deconvolution using end-to-end learning. This means that there is no need to estimate the PSFs and design complicated deconvolution methods. The network should be able to learn to deblur through supervised training on in-focus and out-of-focus image pairs. Zhao et al. [154] uses a residual dense network to estimate the sharp image. However, the defocus images are synthesized using Gaussian kernels, which might not mimic the real-world distortion process. Jiang et al. proposed the DBMID [155] model, which is capable of restoring both defocus and motion-blurred images. They first use a blur-type classification network to predict the type of distortion. Then, the image is restored using a defocus-deblur, a motion-deblur network, or both combined according to the estimated blur type. The defocus training data is captured using z-stack, where the defocus pattern is authentic and better represents real-world defocus blur. In motion deblur, synthetic training data is used. COMI [156] adopts the CycleGAN [159] framework that consists of a deblur module and blur synthesis module. CycleGAN is designed for general-purpose unsupervised domain transfer. The core assumption behind it is that the two-way transformations between the two domains are bijections. However, this assumption does not hold for the blur synthesis module since an in-focus image can correspond to infinitely many out-of-focus ones, depending on the focus level. Wang [157] proposed a multi-scale U-Net [160] that consists of multiple sub-networks that deblur the image at different spatial resolutions. The model first deblur the image at coarser scales, then fuse the deblurred image as well as the feature in the decoder into the encoder of finer sub-networks.

2.3.2 Whole Slide Image Focus Interpolation

Using focus interpolation [161, 162, 163], one can capture only two z-levels and synthesize the middle z-level. Doing the interpolation iteratively allows one to smoothly adjust the focus in a continuous manner, which is usually unachievable in traditional z-stack imaging. It also significantly reduces the scanning time and storage volume since the intermediate z-levels can be interpolated on demand. However, focus interpolation has several drawbacks:

1) Firstly, it can only generate the middle z-level of the two inputs. Generating arbitrary intermediate z-levels requires iterative interpolation, which is not only time-consuming but also accumulates errors, resulting in a large prediction error. 2) Secondly, this method can not extrapolate. The target z-level has to lie within the input z-levels. In practice, this is difficult to achieve since we do not know the exact z-level of the in-focus image. 3) Thirdly, considering the case where we have more than two input images, focus interpolation can only make use of two of them. Although the z-level pairs that are closest to the target are more informative than other combinations, the resting image still contains valuable information, which is not used in this case. 4) Fourthly, this method will not work at all if we only have one input image.

Nicmanis et al. proposed a focus interpolation model [161] that is based on U-Net [160]. The two input images are concatenated along the channel dimension and sent to the network. Different from concatenating the input, DAFNet [162] design two network branches that process the two input individually. The intermediate features are fused using another branch to produce the interpolated image. However, since their goal is only for deblur/autofocusing, they required the two input z-levels to be symmetric to the in-focus z-level, which limits its applicability to arbitrary input z-level combinations. Different from the two interpolation methods above, the method in [163] is capable of both interpolation and extrapolation using a autoencoder-decoder based network. They linearly combine the two latent features extracted by the encoder network, and the combined feature is then feed into the decoder network to produce the interpolated/extrapolated images. By adjusting the two weights of the linear function, they can generate any image lie on this line in the latent space. However, this autoendoer-decoder network is limited in synthesizing realistic high resolution images due to the lack of skip connections. Skip connection is a fundamental architectural design in U-Net that enables the generation of realistic fine details by maintaining the information in the original inputs. But U-Net based architectures are not able to extract intermediate features that encapsulate all information of the input. Consequently they generally can not be used for latent space interpolated/extrapolated.

2.3.3 Whole Slide Image Defocus Synthesis

Although in-focus images are desired for most downstream applications, there is still a lot of quality assurance scenarios that requires out-of-focus images. For example, in training a deep learning based FQA or image deblur model, out-of-focus inputs are needed. Public datasets mainly contains in-focus images and only very few datasets contains defocus ones. Although capturing such defocus images can be achieved thought z-stack, it is both time and storage space consuming. In cases that accessing the tissue slide or microscope is difficult, synthesizing the defocus images becomes a practical solution. In the WSI FQA and deblur literature, a lot of works have incorporated synthesized defocus images in their model design or training. Existing methods synthesize the defocus effect by convolving the in-focus image with blur kernels of various kind: Gaussian, Disk, or parametric PSF. More information is shown in Fig. 2.2.

Table 2.2: Whole Slide Image FQA and deblur methods that use synthetic defocus training data.

Task	Gaussian Kernel	Disk Kernel	Estimated Parametric PSF
WSI FQA	[38, 52, 61]	[61]	[59]
WSI Deblur	[154]	N.A.	[153]

However, synthesizing defocus through convolution has limitations. For the Gaussian and disk kernel, they barely mimic the real world out-of-focus process. Although the estimated PSF is closer to the real distortion process, it requires the optical specifications of the microscope and the tissue, which is tedious and challenge to collect for each microscope at each focus distance. However, the study of synthesizing real-world defocus blur in WSI is still vacant. Nevertheless, although not designed for defocus synthesis, some focus interpolation methods [161, 163] are able to produce more realistic distortions than the traditional convolutional-based methods. But these interpolation methods also suffers from the disadvantages as described in Sec 2.3.2. The major drawback is that they requires two input images. For defocus synthesis purpose, at least one of the input should be out-of-focus, which is controversial to the goal.

Chapter 3

High-Efficiency Focus Quality Assessment for Whole Slide Image

3.1 Introduction

Pathology is the study and diagnosis of disease, which involves the examination of surgically removed organs, tissues, or bodily fluids. Subfields of pathology include surgical pathology, cytopathology, molecular pathology, etc. Surgical pathology examines surgically removed tissues with the naked eye or under a microscope (histology). Cytopathology studies diseases on the cellular level, which involves the examination of free cells from bodily tissues or fluids. Molecular pathology studies diseases on the molecular level. Both histology and cytopathology require the specimen to be processed to satisfy the requirements of optical brightfield microscopic viewing. No matter what kind of pathology, a professional pathologist is required to be present to conduct the examination. In areas that do not have adequate specialists, the prepared samples need to be physically delivered to other locations. To fulfill the lack of pathologists and enhance the efficiency of the diagnosis workflow, digital pathology emerged during the 1960s. In digital pathology, glass slides are scanned into digital images that can be viewed, stored, shared and analyzed on computer systems. Digital pathology addresses the problem of uneven distribution of pathology services, making pathologists able to work from any place that has an internet connection.

It not only makes telepathology possible but also facilitates the development of automated image analysis systems and lowers the cost of medical education.

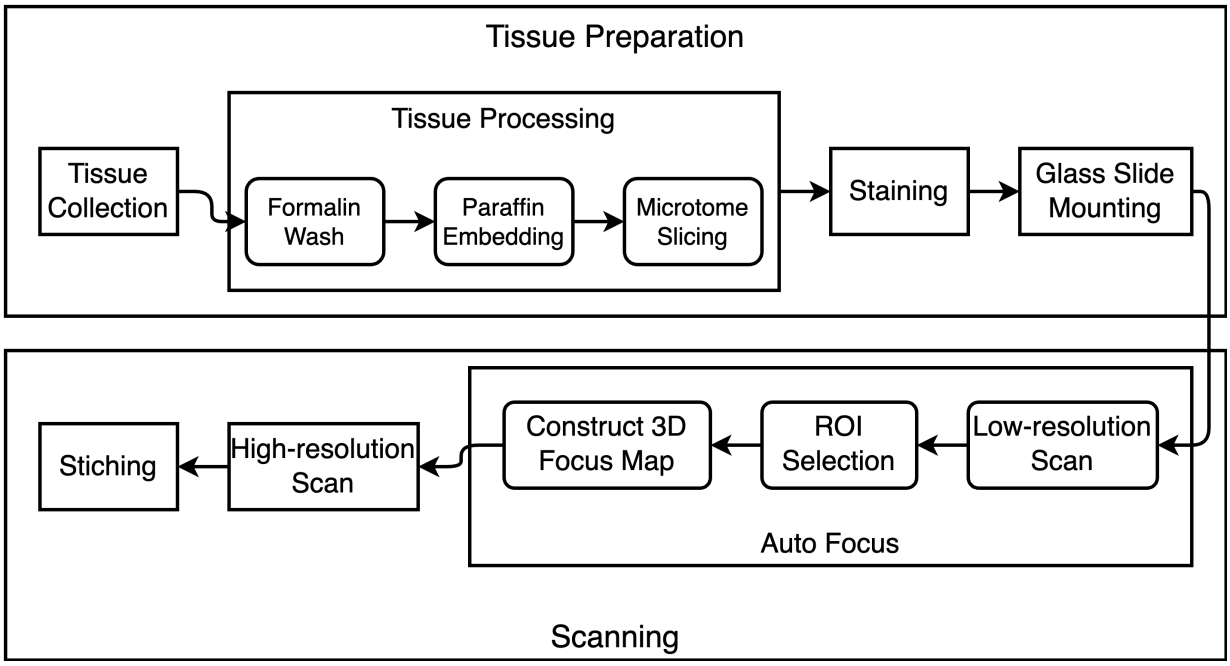


Figure 3.1: The tissue preparation and scanning pipeline for digital pathology.

The tissue preparation pipeline is depicted in the upper part of Fig 3.1, it usually includes tissue collection, tissue processing, tissue staining, and glass slide mounting. In tissue processing, the biological samples to be examined undergo formalin wash and paraffin embedding, which is referred to as Formalin-Fixed Paraffin-Embedded (FFPE). The fixed biospecimen is then sliced using a microtome (sectioning). The objective is to get a thin and transparent slice in which only a single layer of cells is present. This is a basic requirement for an optical brightfield microscope. In the staining stage, depending on the type of organ, different stain compounds will be applied to the sample to enhance the visibility of structural details. In the glass slide mounting step, a coverslip is placed over the tissue to provide protection from environmental contamination and damage. The final stage is scanning, where the slide is scanned using a WSI scanner. As depicted in Fig 3.2, the major components of WSI scanner's hardware includes a brightfield light source, a condenser, an

automated stage, an objective lens, a tube lens, a camera sensor, and a built-in image processor. While the objective lens, tube lens and sensor are fixed, the automated stage can move freely, allowing the scanner to capture different locations of the slide and set different focal planes for each individual tile. An objective lens with a high NA is typically used in WSI systems in order to achieve a higher resolution. Because of this, the DOF of the WSI system is usually on the micron level, which makes precise focusing during the scanning procedure difficult. The NA is defined as

$$NA = n \times sin(\alpha) \tag{3.1}$$

where n is the refractive index of the medium between the objective lens and the sample. For the dry lenses that use air as the medium, $n = 1$. For the oil immersion lenses, $n = 1.52$. As illustrated in Fig 3.2, α is the half-angle of the cone of light entering the objective lens [164]. The optical resolution R of a microscope is defined as the minimum distance between two airy disks that can be distinguished on the image plane, which is defined as

$$R = \frac{0.61\lambda}{NA} \tag{3.2}$$

where λ is the wavelength of the light source. It is straightforward to see that to maximize the resolving power of a microscope, we need to lower R, which means to either use a light source of shorter wavelength λ or increase the NA. For brightfield microscopes, λ is lower bounded since the visible light is used. The only choice is to use a lens with larger NA. However, this comes at a cost of shallow DOF. Due to the non-zero resolution R, there is a range around the focus plane in which image sharpness does not change [164]. This range is often referred to as DOF. Objects that lie within the DOF will stay sharp while those that lie outside of DOF will be out-of-focus. DOF can be formulated as

$$DOF = \frac{1.22\lambda}{NA \times tan(\alpha)}. \tag{3.3}$$

It is easy to see that the DOF is proportional to the resolution R, meaning most microscopes

will have a very shallow DOF. This makes the autofocus a very challenging task in the WSI system.

Different from the optical resolution, the resolution of the entire system, including optics and digital sensors, is defined as the size of the pixel in the final image. The maximum resolution of the system usually ranges from 0.1 $\mu m/pixel$ to 0.3 $\mu m/pixel$ [74]. A general scanning pipeline is also illustrated in the bottom part of Fig 3.1. The first step is to generate a 3D autofocus map. The camera first captures a low-magnification preview of the slide. Using the preview as a guide, the tissue detection algorithm finds the location(s) of the tissue and this process is known as Region of Interest (ROI) selection. Finally, the tissue is separated into several tiles or strips based on the scanner type. The optimal focal plane for each tile or strip is then determined by an autofocus algorithm. Many scanners place focus points on every "nth" tile or strip in order to save time because of WSI's high resolution. A 3D focus map is then obtained by stitching (and interpolating) these focal planes. In the second step, an objective lens with higher magnification (20X or 40X) is used to scan each tile/strip at a time, according to the 3D focus map generated in the last step. This is achieved by moving the automated stage along the optical axis (z-axis), a process that is frequently referred to as sample holder scanning, as opposed to objective lens scanning or camera sensor scanning. The final step is to stitch the scanned tiles/strips together into a single WSI. Certain post-processing techniques, like color calibration and compression, may also be applicable afterward.

Any issue that arises during the slide preparation process may degrade the quality of the samples that are prepared, which may affect the accuracy of the diagnosis taken by the pathologist or automated systems. The most common artifacts in the prepared slides are foreign objects, uneven tissue terrain, tissue folds, pen marks, air bubbles, dust in the slide, stain variation, etc [165]. Since they cause the slide's thickness to vary and cause the autofocus system to malfunction, the majority of these physical artifacts will show up in the scanned image as different kinds of out-of-focus blur. Due to the microscope's shallow DOF, the out-of-focus artifact is usually quite noticeable.

The physical artifacts in the prepared tissue can be categorized into three types. The first kind of artifacts cause variations in thickness in the prepared specimen can significantly affect the focus. The first kind of artifact introduced during slide preparation changes the

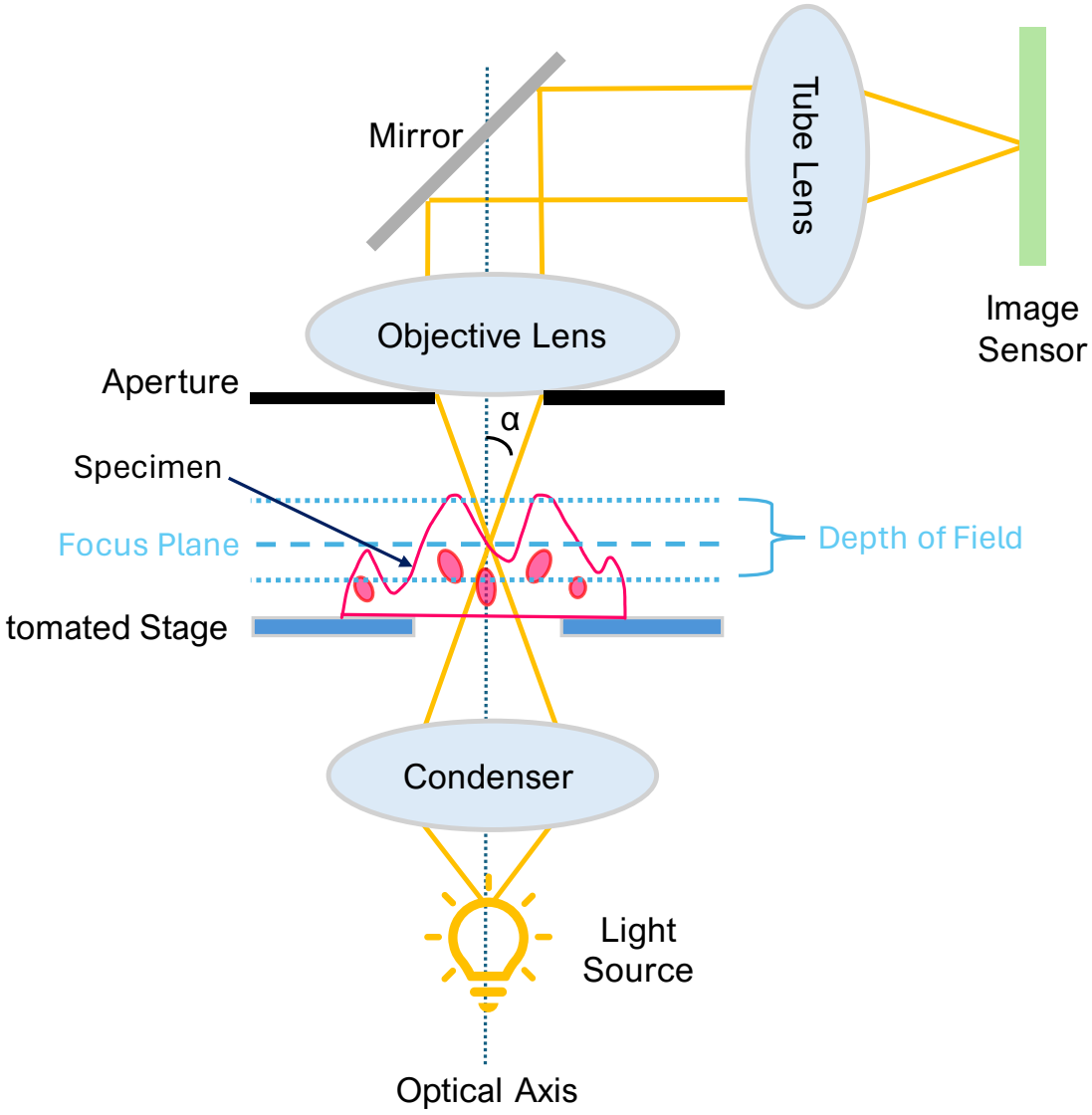


Figure 3.2: The general structure of a WSI scanner. α is the half-angle of the cone of light entering the objective lens.

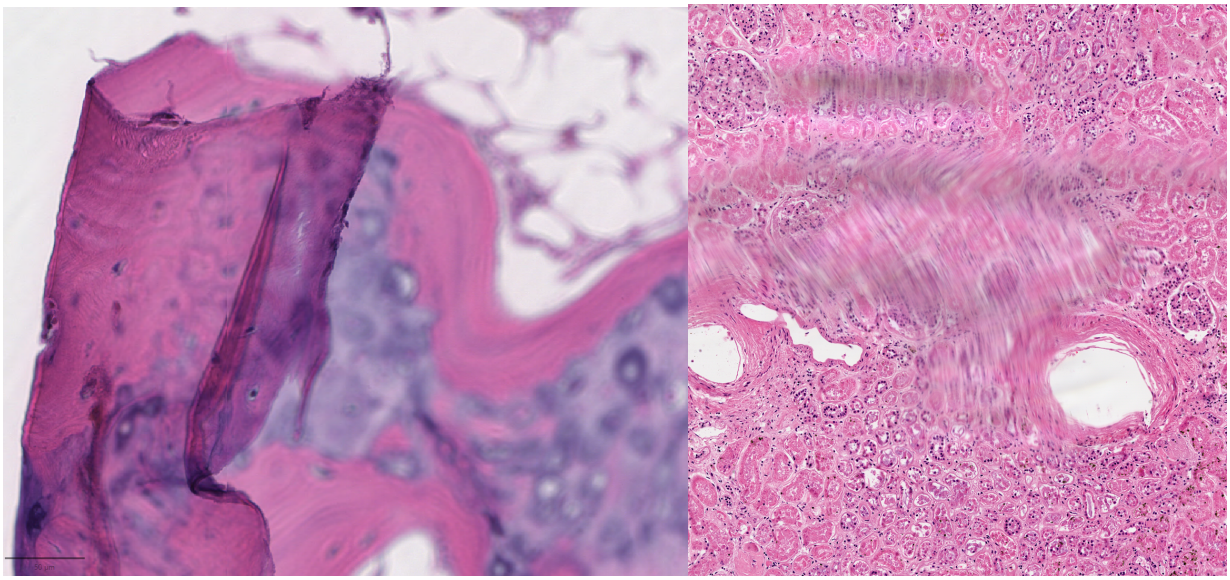
height of the tissue, making the surface uneven. These artifacts include tissue folding and tearing. Due to the shallow DOF of microscopic lenses, it is challenging for the autofocus

system to make sure every tile is in focus using a limited number of focus points. Some artifacts, such as air bubbles and wrong coverslip placement, disrupt the optic path. Most WSI platforms require the medium between the tissue and the objective lens to be either air, water, or oil. The reflective index of the medium is an important preset parameter in the system, which determines the NA, optical resolution and DOF. The presence of air bubbles and wrong coverslip placement make the reflective index deviate from the design value, resulting in malfunctions of the autofocus system. Additionally, the complexity of biological specimens, which may have regions of differing optical properties such as refractive index, further complicates the autofocus process. The third kind of artifact impacts the brightness and contrast of the tissue. These artifacts include marker strokes and wrong staining procedures. Markers are frequently used in clinical applications to highlight areas of interest on the coverslip. The areas under the marker strokes tend to be much darker and have a lower contrast. While staining is used to enhance the contrast and visibility of specific tissue components. For example, IHC uses antibodies tagged with dyes to visualize the presence and distribution of specific proteins. By stating the tissue incorrectly, the contrast of the slide can be low. Due to the low contrast, both marker strokes and wrong staining procedures can contribute to the malfunctions of the autofocus system. These challenges necessitate robust FQA techniques to ensure the reliability of WSIs for diagnostic purposes.

Some examples are shown in Fig 3.3. Even if the tissue is well-prepared, a lot of errors in the scanning process can lead to focus miss alignment which results in out-of-focus blur. For example, errors in the autofocus algorithm, an insufficient number of focus points, thermal issues of the mechanical system and vibrations can all lead to an incorrect 3D focus map, which represents the tissue terrain and is used to set the focus plane of each tile/strip.

In traditional microscopy where no digitization is introduced, this phenomenon is alleviated to some extent since pathologists evaluate multiple focus planes. Despite the fact that certain scanners can take pictures in multiple focal planes (z-stacking), this requires a corresponding increase in scanning time and storage. Consequently, z-stacking has been largely confined to research [166]. Although z-stacking seems to be the answer to fix out-of-focus blur, most of these physical artifacts cannot be simply removed from the image

by varying the focus plane. For instance, as shown in Fig 3.3, the other layers in a folded tissue will always be visible above or below the layer you choose to focus on, resulting in a blurring image.



(a) Tissue folding (b) Air bubble

Figure 3.3: (a) Folding artifact on the WSI of bone [3]. The tissue's top layer appears to be in-focus, while the bottom layer is out-of-focus. The bottom layer still obstructs the upper layer's perceptual clarity even though the upper layer is in-focus. (b) Air bubble artifact caused by incorrect coverslipping. Because of the air bubble's distinct refractive index, light is diffracted differently from the other parts, resulting in a blurry image.

The problem of out-of-focus artifacts in digital pathology is a huge bottleneck in existing high throughput WSI scanning platforms, making them difficult to integrate into clinical workflows. WSI scans are required to be manually inspected for FQA on the pixel-level, which is (a) highly tedious and time-consuming; and (b) subjective to an individual scoring that often causes inter/intra-variability issues. Consequently, an objective FQA plays an essential role in the quality control pipeline. The functions of an objective FQA algorithm include a) determining whether a slide needs to be rescanned locally or globally; and b) providing an FQA map that can be used for visual inspection and validation; and c) guiding

the autofocus algorithm [167, 168].

Both knowledge-driven and data-driven FQA models have been developed in the literature.

Data-driven Focus Quality Assessment. Recent developments involve supervised training of CNNs on the image patch labels of a given focus dataset of WSIs, where the network is either adopted from a pre-designed architecture followed by some minor adjustments [52, 61] or tailored from scratch [59, 1, 169]. The selection of training datasets can also be divided into two categories of either synthetically generating out-of-focus (defocus) images by convolving in-focus patches with artificial blur kernel with different grades (i.e. classes) [59, 52, 61], or scanning thee prepared slides in different focal planes (z-levels) to generate real blur classes [1]. Existing open source software solutions such as CellProfiler [170] and HistoQC [171] adopt variants of such models for FQA of WSIs. The high computational complexity and the lack of transferability are the main drawbacks of these models.

Knowledge-based Focus Quality Assessment. Numerous methods have been developed in the literature based on a wide variety of domain knowledge, including human visual system models and microscopic optics models [57]. Although these methods may have lower computational costs compared to data-driven ones, their accuracies are relatively low compared to data-driven solutions, as will be shown later.

General Purpose Image Sharpness Assessment. FQA and ISA are both subdomains of IQA with similar objectives but different scopes of application. FQA is developed specifically for digital pathology, with a focus on high precision and time efficiency to ensure that the focus quality of microscopic images is sufficient for precise diagnosis. FQA models often make use of domain-specific information relevant to pathological analysis or microscopic imaging. On the other hand, ISA is more widely applicable in a variety of scenarios, including general photography, industrial process imaging, surveillance, etc. Usually, it assesses the overall perceived sharpness of images without taking domain-specific factors into account [93, 107, 116, 108, 58]. Nevertheless, some ISA models are also applicable in the case of digital pathology [2].

How Existing Models are Limited? Despite great performances of data-driven

approaches such as CNN in deep learning [52, 61], they have not been integrated into high throughput scanners for quality control purposes due to two main reasons. First, the computational complexity of data-driven solutions is often too high to process GigaByte WSIs. We explain this in the example as follows. Despite the FQA models taking a few seconds to process one patch from WSI that are fast enough, the story is quite different for high-throughput scanning systems. Depending on the vendors, several hundreds of glass slides can be mounted in scanners (e.g. Philips Ultra Fast Scanner accepts 300 slides of 1"x3" and Huron TissueScope-iQ accepts 400). In clinical settings, all scans should be completed during the night hours (less than 12 hours time frame) to be ready for diagnosis for the next day. Each slide is usually scanned at 0.5um/pixel @20X magnification, containing $\sim$ 1cm $\times$ 1cm tissue which translates to $25,000 \times 25,000$ digital WSI, yielding $\sim 2,500$ patches of 1024×1024 (50% overlap). Assuming two models are used for assessment, i.e. M1: DenseNet-10 and M2: FocusLiteNN (our proposed model), the time taken for two models to complete the task is

$$M1: \ 2,500(\text{patches/WSI}) \times 300(\text{WSI}) \times 0.355\text{sec/patch} = 73.96 \text{ hour}$$
$$M2: \ 2,500(\text{patches/WSI}) \times 300(\text{WSI}) \times 0.017\text{sec/patch} = 3.54 \text{ hour}$$

Clearly, the speed gain from model M2 over M1 is obvious. The limitation in computational resources becomes equally important as the precision when choosing FQA models for GigaByte WSI processing [172, 173]. The second limitation is the lack of transferability of CNNs which becomes a barrier to process WSIs across different tissue stains and scanner variations.

Contributions. Our aim in this paper is to address the challenges in data-driven FQAs. In particular, (a) we build a highly efficient extremely lightweight CNN-based model i.e. FocusLightNN [1] that maintains fast computations similar to the knowledge-driven methods without excessive hardware requirements such as GPUs. The database used for training plays a crucial role, for which we suggest a training dataset using FocusPath [2] which encompasses diverse slides across nine different stain colors. We hypothesize that the stain diversity greatly helps the model to learn diverse color spectrums and tissue

[1]Codes and models are available at https://github.com/icbcbicc/FocusLiteNN

structures. (b) For algorithm evaluation and comparison, we introduce a novel comprehensive evaluation dataset that is annotated and compiled from the TCGA repository. Comprehensive experiments and analyses are conducted that demonstrate the superior precision-speed compromise of the proposed approach.

3.2 FocusLiteNN Model

3.2.1 Difference in Natual Image and WSI

In clinical applications, the focus quality of an image directly impacts the ability of pathologists to resolve biological structures in detail, potentially affecting diagnostic decisions. Although FQA seems to be a special case of the more general task ISA, digital pathology images differ from general photographic images in several ways. As a result, general ISA models, while effective in many general photographic scenarios, are often insufficient for the unique demands of digital pathology.

Notably, the first difference is the object being captured. Most WSIs are shift and rotation-invariant and can have varying contrast, colors, and textures depending on the staining process and tissue type. In natural images, objects can vary significantly in scale. Some objects may be up-close and detailed, while others are distant and less defined. However, the scale of structures in WSI tends to be more uniform. Fig 3.4 shows a comparison of the distributions of image gradients, normalized by average luminance, between in-focus natural images and in-focus WSIs. Natural images have a broader distribution of gradients, implying a wider range of edge sharpness and contrast. In contrast, the distribution for WSIs is narrower and peakier, suggesting a more uniform level of edge sharpness. The figure suggests that WSIs tend to have a more consistent level of detail. While some ISA methods rely on semantic information of natural images [120, 121], the information in WSIs is very different and the prior of natural images can not be easily transferred to FQA. Photographic images also undergo different distortions before, during, and after image capture compared to WSIs. For example, atmospheric turbulence, moving objects, shaking camera, out-of-focus and lens aberration can all contribute to the blur in

photographic images. While in WSIs, out-of-focus and lens aberration are the main cause of blur.

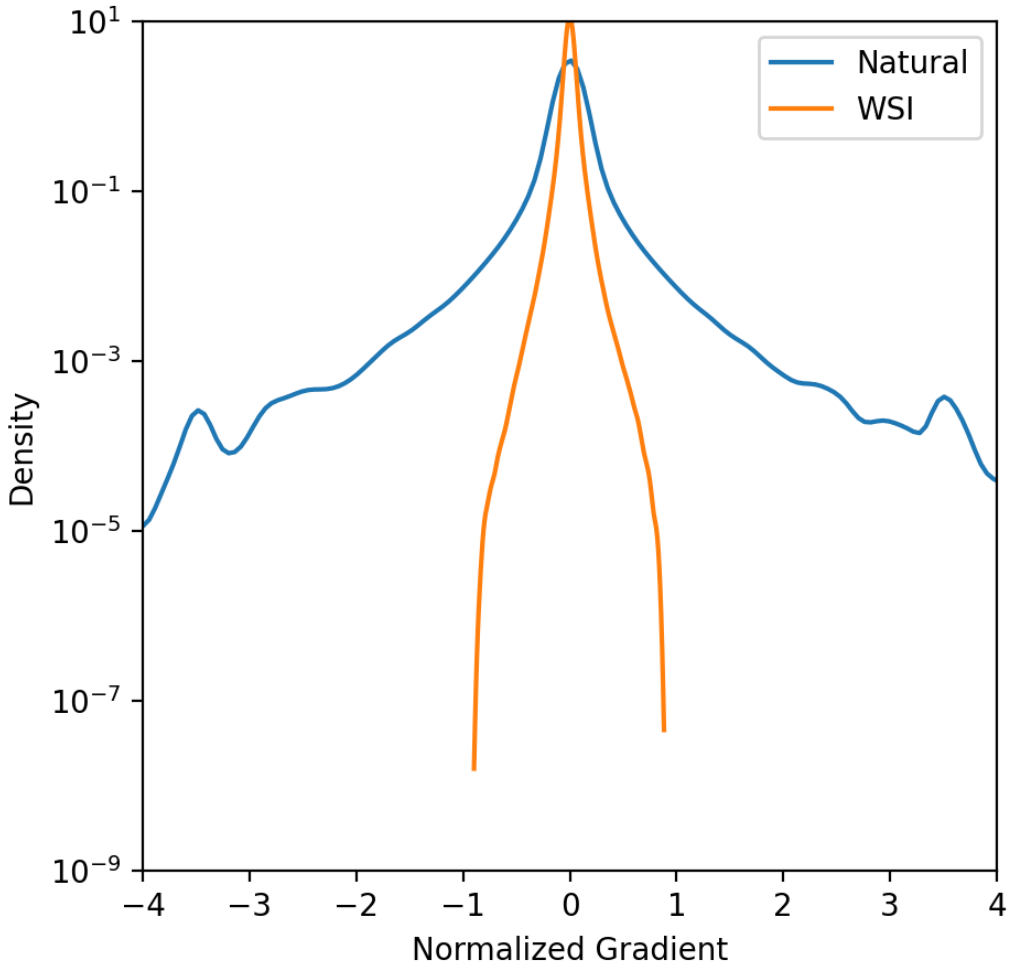


Figure 3.4: The distribution of the image gradient normalized by the average luminance. The distribution of sharp natural images is different from the one of sharp WSIs.

The second difference lies in the imaging device, such as the optical system, illumination conditions, sensors, and image processing algorithms. The optical design of the

microscopes in WSI scanners differs substantially from general cameras. Microscopic optics are specifically designed to resolve microscopic details at the cellular or even molecular level that are not visible to the naked eye, requiring lenses that can achieve high resolutions with minimal aberrations. This is usually achieved through using a lens with a high magnification rate, high NA and medium with a higher reflective index than air. For example, the NA for a "large" aperture (F1.8) photography lens is only 0.27. However, common NA for the 40× microscopic lens used in digital pathology ranges from 0.75 to 1.2 [74]. As a result, the two optic systems often have substantially different spherical abbreviation patterns. The most noticeable difference is that the microscopic lens usually has a very shallow DOF, which means that only a very thin layer of the specimen is in focus. However, most consumer camera lenses have a relatively deep DOF to make sure the object is in focus and sharp. Some professional macro lenses may have a 1× magnification ratio, which makes the DOF a little bit shallower. However, it is still not comparable to microscopic lenses.

The illumination condition is also different. The microscopes have controlled lighting conditions, often through transmitted light from below the sample. This controlled lighting system provides constant and even illumination across the entire sample, which is crucial for capturing the details in specimens without interference from outside illumination. On the other hand, consumer cameras rely on ambient light or flashlights which are more variable and less controlled, which may impact the observation of sharpness and contrast.

The image sensor and image processing algorithms used are different. WSI scanners use sensors that are optimized for color accuracy and resolution to account for the fidelity of stains and detailed structures necessary for diagnostic purposes. The scanned images often have different black-level, pixel gain, photon noise and dynamic range compared to general image sensors [59].

In conclusion, due to the inherent differences between the content and imaging devices, the appearance and distortion types of WSI and photographic images are different. As a result, ISA models typically do not account for the specific characteristics of WSIs. Consequently, these general models can struggle to distinguish between genuinely out-of-focus areas and intricate details of the tissue structures, leading to inaccurate assessments of

image quality. Existing FQA models in digital pathology aim to address these challenges by incorporating domain-specific knowledge, such as the optical characteristics of microscopes [57]. However, existing FQA models, whether knowledge-based or data-driven, have inherent limitations that can affect their performance and applicability. Knowledge-based models, which rely on handcrafted features to assess image quality, are usually suboptimal. Such models often rely on oversimplified assumptions about image characteristics that are supposed to indicate focus quality, potentially missing important features that differentiate between in-focus and out-of-focus images. On the other hand, data-driven models, particularly those leveraging deep learning techniques, offer the advantage of learning from vast labeled data to identify the features that are important for FQA. These models usually have better performance compared to knowledge-based ones. However, this performance gain comes at the cost of being computationally expensive. These CNN-based FQA models usually require graphics processing units (GPU) to conduct the inference. Nevertheless, due to the high-resolution requirement, the size of WSIs are much higher than normal photographic images. Considering the high computational cost and large data volume, computational efficiency becomes a bottleneck that significantly hinders the workflow in WSI platforms, where timely scanning is critical for subsequent diagnosis.

3.2.2 Assumptions

Our proposed model, FocusLiteNN, is designed to overcome the limitations of existing FQA methods by significantly enhancing time efficiency without compromising accuracy. In order to benefit from the deep learning framework that can learn from data, we build FocusLiteNN as a CNN. The main idea of our model design is to prune the complicated network architectures of existing CNN-based FQA and ISA models. Unlike most CNNs that use a hierarchical structure of convolutional layers, FocusLiteNN consists of only one convolution layer. Such a shallow design is based on prior knowledge about the WSIs and a deep understanding of the scanning workflow. Consequently, out-of-focus blur in WSIs can be characterized using a relatively simple model. We summarize the knowledge and understanding into the following assumptions: (a) The scale of structures in WSIs are relatively uniform, and (b) the distortion process is relatively easy to characterize, and (c)

sharpness information is mainly encoded in the low-level information rather than high-level information, and (d) color information is important to distinguish the contrast in stained tissue. We further describe these assumptions in detail as follows.

Uniform Scale of Structures The first assumption suggests that the structures within WSIs are relatively uniform in scale compared to natural images. Biological tissues and cells have inherent size ranges that do not vary as widely as objects in natural images. The objective lenses used in microscopy are selected based on the size of the structures being studied. The most commonly used magnification rates are $20\times$ and $40\times$ in pathology applications. When features in an image maintain a relatively consistent scale, it eliminates the need for complex network designs. For example, hierarchical convolutional layers, convolutional layers with different kernel sizes, convolutional layers with strides larger than one, dilated convolutional layers, and pooling layers are all used to capture and aggregate features of different scales. These complex designs do not provide significant benefits while greatly boosting the computational cost.

Controlled Distortion Process The second assumption indicates that the distortion process causing out-of-focus blur in WSIs is relatively easy to characterize compared to natural images. This is mainly due to the image capturing taking place in a well-controlled environment, which is the WSI scanner. Wrong focus plane and spherical aberration are the two major causes of blur in WSI. According to Weber's law, the uniform illumination within the scanner also guarantees the perceptual contrast will not change due to brightness variations. As shown in Fig 3.5, the contrast of the underexposed patch (orange boundary) is very low, and the textures within it are not visible. However, we know that this area is full of textures similar to the patch with blur boundaries. The image sensor in WSI scanners also has a higher Signal to Noise Ratio (SNR) compared to consumer cameras. The compression algorithms used in WSI scanners are either lossless, such as JPEG2000, or high-quality lossy. However, in addition to wrong focus plane and spherical aberration, atmospheric turbulence, moving objects and shaking cameras can also contribute to blur in general photography. The spherical aberration is also more noticeable for consumer camera lenses. The lower SNR of the sensor and high compression ratio in post-processing

[2]The image is taken from https://unsplash.com/photos/opened-door-kmY-rsl7BRw, which is free to use under the Unsplash License

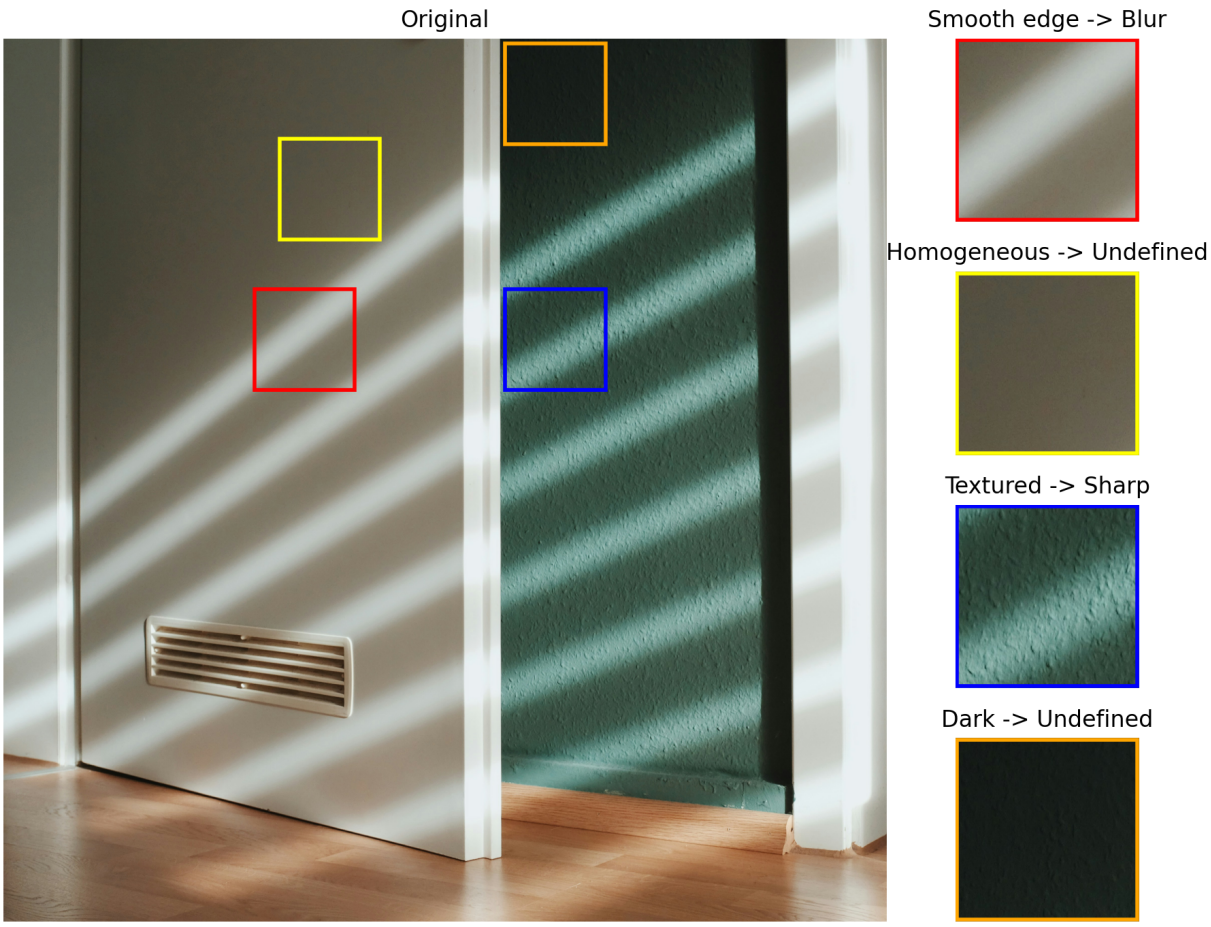


Figure 3.5: This figure [2]illustrates how semantic information affects the perceived sharpness of natural images. All parts of the original image are in focus. However, if individual local patches are isolated from the whole image, their perceived sharpness might be different from the whole image. Textured patches are sharp and patches containing smooth edges are blurry. The sharpness of smooth, constant, overexposed, or underexposed patches is undefined.

further degrade the quality of general photography images. As a result, the sharpness of general photography images is more difficult to characterize due to the complex distortion process, which requires a more complex network design. On the other hand, a distortion

process that is easy to characterize allows for simpler models.

Sharpness Encoded in Low-Level Features The third assumption is that sharpness information is primarily contained in the low-level features of an WSI rather than in the high-level features. However, this might not hold true for natural images. The local sharpness of natural images is correlated with high-level semantic information [120]. We illustrate this in Fig 3.5, which is captured in an office with a high DOF lens. The focus is accurate and all objects in the room appear sharp. However, if we take a closer look at a homogeneous patch (yellow boundary) cropped from the clear white door, we will find it difficult to assess the sharpness of the patch due to the lack of texture. Even with the presence of edges, the sharpness of a local patch might still differ from the whole image. For example, some objects might cast a shadow on the same door, which turns out as edges. However, due to the scattering light, the edge of the shadow might be soft, meaning a larger spread over the edges. A lot of edge-based ISA might consider this patch (red boundary) as blurry. These examples show that semantic information is crucial for ISA. Capturing such information needs a deeper and more complex CNN. However, due to the unique characteristics of biological samples, the sharpness of WSI can be solely determined through low-level features without considering semantic information. Examples are shown in Fig 3.6. The presence of smooth edges/textures (red boundary) means that the area is out-of-focus. Homogeneous patches (yellow boundary) indicate no tissue is present in that area, which can be considered as out-of-focus. Low-level features in images include edges, textures, and basic shapes that are extracted by early layers in a CNN. This assumption supports the idea that a single convolutional layer can effectively capture the necessary details for FQA in WSI.

Importance of Color Information The fourth assumption recognizes the importance of color information in distinguishing contrast within stained tissue samples. In the context of digital pathology, different stains are used to highlight various tissue components or proteins. The contrast created by these stains is vital for assessing tissue structures. In certain stains and tissue types, the contrast is presented as differences in hue and saturation, instead of luminance. Assessing the sharpness using the luminance channel (grayscale image) alone will be less accurate. In addition, by considering color information, the model might further capture chromatic aberration and take it into account in sharpness

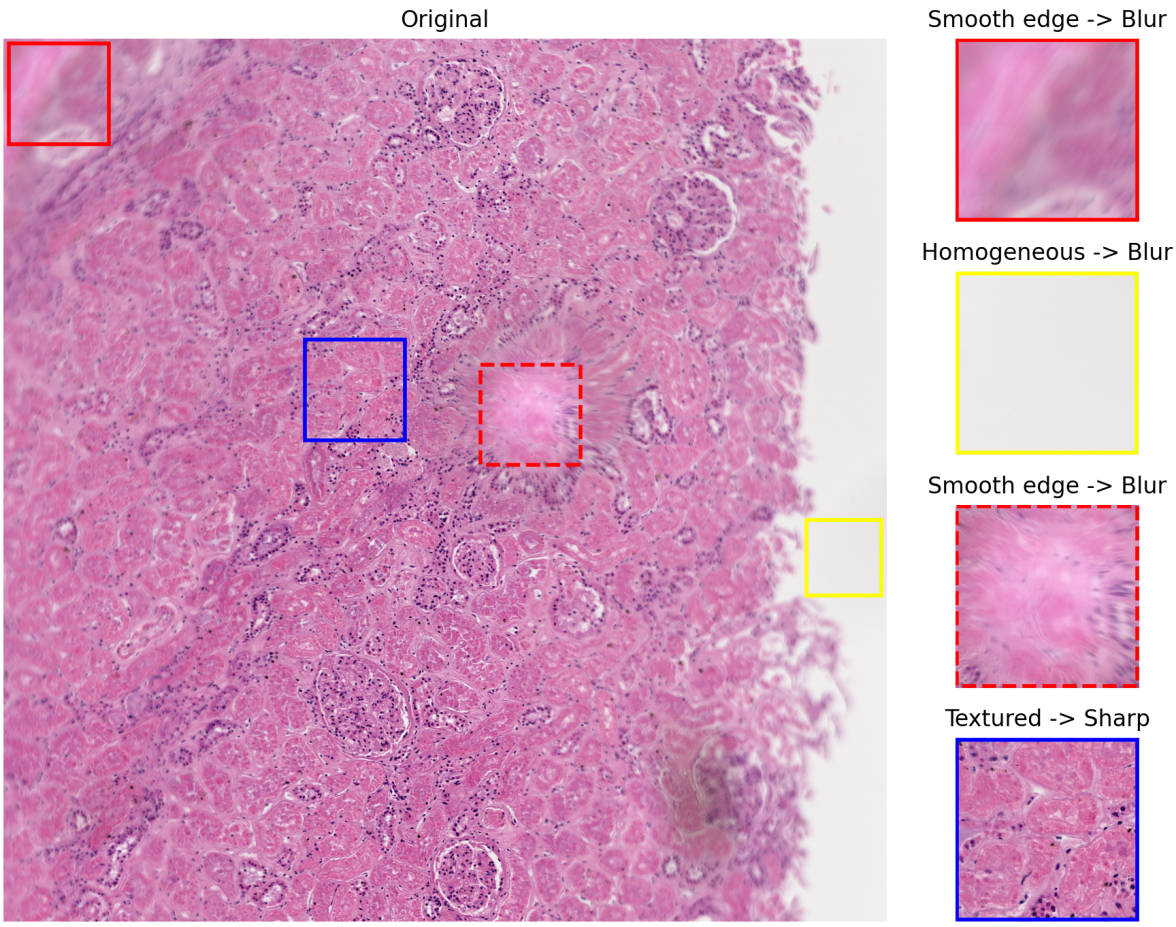


Figure 3.6: This figure illustrates that the perceived sharpness of WSI is irrelevant to the semantic information.

assessment. Chromatic aberration is caused by the lens's failure to focus light of different wavelengths into the same point. The reason behind this is that the refractive index varies with the wavelength of light, which means the focal length is also different for different wavelengths. The chromatic aberration provides an important cue for the focus distance, which contributes directly to the level of blur of the image. We show this phenomenon in Fig 3.7. Fig 3.7 (a) shows the mechanism of the chromatic aberration. It is easy to find that different focus distances result in different aberration patterns. In Fig 3.7 (b),

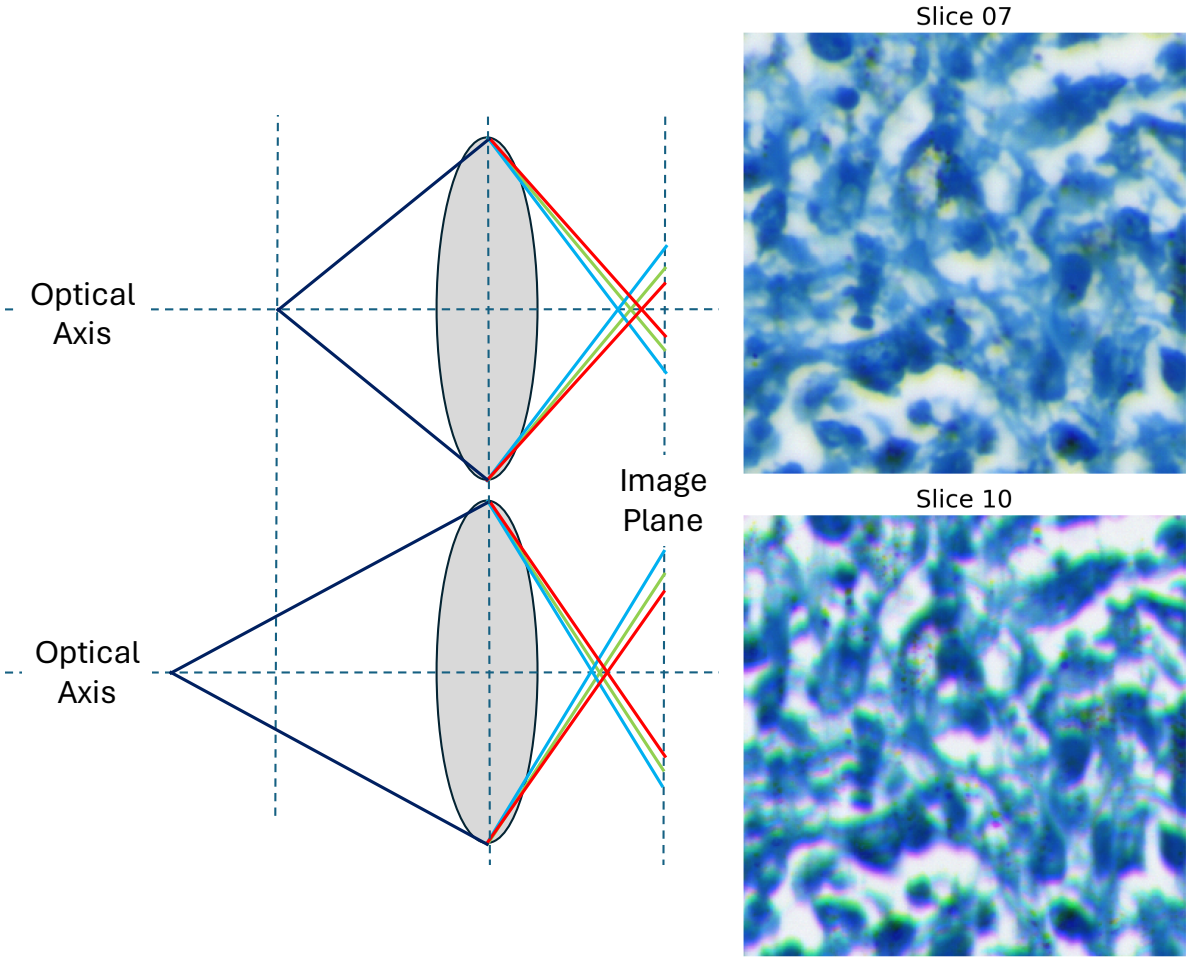
example images captured at different focus distances are shown. Note that the chromatic aberration is different in these two images: the upper one has some green/yellow color along the edges, while the lower one is red/blue. To summarize, these two observations suggest that a model designed for FQA in WSIs should be sensitive to color information.

In conclusion, FocusLiteNN is designed to efficiently address the FQA challenge by leveraging these four assumptions. By acknowledging the uniformity of feature scales, the simplicity of distortion, the localization of sharpness information in low-level details, and the importance of color, the model can operate with a single convolutional layer. This simple approach aims to maintain high accuracy while improving time efficiency in the high-throughput scanning workflow.

3.2.3 Model Design and Analysis

Patch Size Due to the uneven height of tissue in a slide, the focus quality varies across an WSI. This requires FQA models assessing local focus quality. This is usually achieved by separating the WSI into many patches and predicting the sharpness for each of them. Using a patch-based approach, we assume that the sharpness level is uniform within that patch. Choosing a reasonable patch size is important to capture biological structural information while avoiding the ambiguity issue.

The disadvantage of using a small patch size for FQA is that there is not sufficient structural information within such a small area to reliably differentiate between in-focus and out-of-focus. Moreover, the lack of detail combined with the small patch size can result in sensor noise being the dominant source of high-frequency components, which further complicates the task. This phenomenon is illustrated in Fig 3.8 where two in-focus patches are extracted. The one with the yellow-green boundary is 64×64 while the yellow one is 235×235. The physical size of a 64×64 patch is around $16\mu m \times 16\mu m$ in this figure. Since the normal range of size of nuclei is around $5\mu m \sim 20\mu m$, it is very likely that a $16\mu m \times 16\mu m$ will not contain any nucleus. For example, the yellow-green patch does not contain any noticeable biological structure information, such as the nucleus or the boundary of a cell. The color of the nuclei is usually darker than other regions due to staining, which provides enhanced contrast. As a result, the yellow-green patch has less



(a) Systematic view of the chromatic aberration at different focus distance

(b) Examples of the chromatic aberration shown at different focus distance

Figure 3.7: (a) The systematic view of the chromatic aberration. Different focus distances result in different aberration patterns. (b) Example images captured at different focus distances. Note that the chromatic aberration is different in these two images: the upper one has some green/yellow color along the edges, while the lower one is red/blue.

structure and has a lower contrast compared to larger patches which contain biological structures (yellow one). Patch size not only affects the perceived sharpness of in-focus

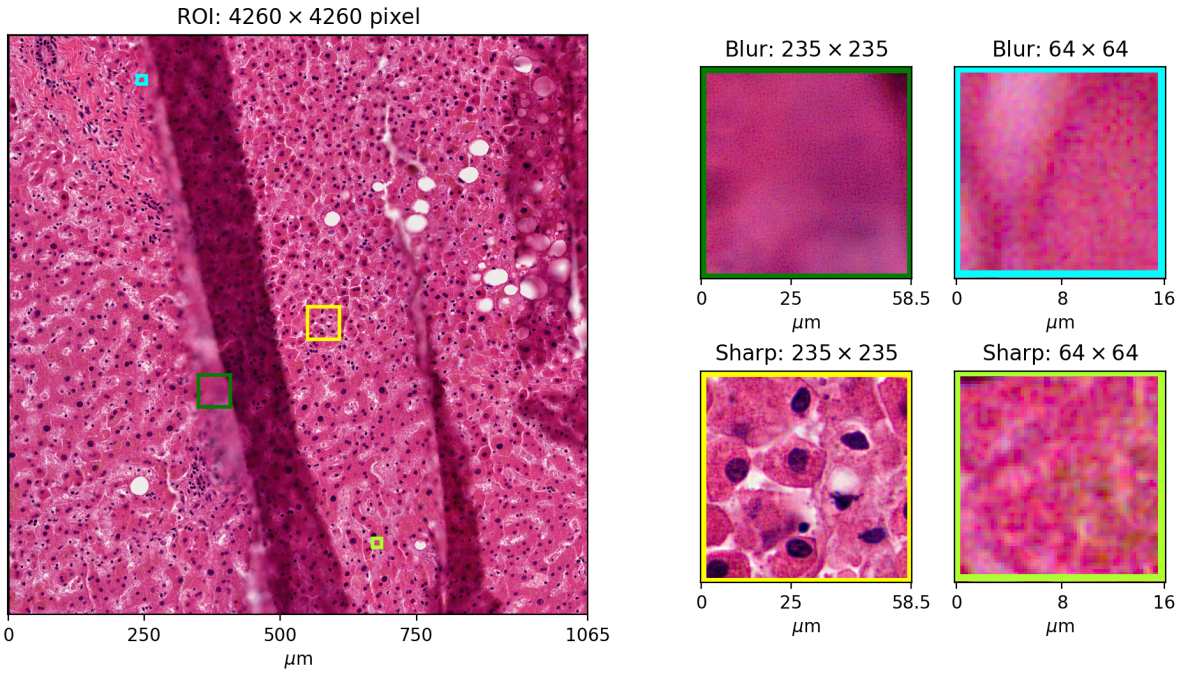


Figure 3.8: This figure illustrates the influence of patch size on sharpness assessment. While both the yellow (235×235) and yellow-green (64×64) patches are in-focus, the yellow one looks sharper since it captures more biological structural information, such as the nucleus. While both the cyan (64×64) and green (235×235) patches are in the same level of out-of-focus, the cyan one looks sharper since the sensor noise is more pronounced.

regions but also out-of-focus areas. Two out-of-focus patches are also extracted, one is 64×64 (cyan) and the other is 235×235 (green). In the cyan patch, sensor noise pops out due to the blurry background, which contributes to the high-frequency components. As a result, The perceptual sharpness of the cyan patch is higher than the green one due to the larger portion of high-frequency components. In conclusion, choosing an oversmall patch size will result in a miss in structural information and a boost in perceived noise, which ultimately makes FQA inaccurate. Using an overlarge patch size also has problems. The first problem is that both in-focus and out-of-focus regions can appear in the same large patch, which makes the prediction ambiguous. The second problem is the increase

in computational complexity.

Based on our empirical observation and quantitative analysis, we find the patch size of 235×235 works the best for WSIs with $40\times$ magnification, which is the most popular setup for the purpose of pathological diagnosis.

Model Architecture In the FQA literature, most knowledge-based models adopt relatively simple features, such as edges and high-frequency components. These features usually can be approximated by convolving the original image with a set of kernels. For example, commonly used edge detection kernels include Canny, Sobel, Prewitt, Scharr, Laplacian, DoG, LoG, etc. Transforming images from the spatial domain to the frequency domain can also be achieved by convolving the image with FFT and DWT kernels. This demonstrates the power of using simple filters without complex hierarchical architecture to extract sharpness-related features. However, most of these handcrafted filters are designed for general ISA which may not account for the specific characteristics of WSIs. As a result, these handcrafted filters may be suboptimal for the task of FQA. Recent advances in deep learning have demonstrated the power of CNN being able to extract task-specific features through learning the convolutional filters from data. The multiple convolutional layer design in deep CNNs is responsible for capturing multi-scale and semantic features. However, these complex designs increase the computational complexity while do not provide significant performance improvement. Based on our assumptions that 1) WSIs are relatively uniform in feature scales, 2) the distortion process is controlled, and 3) the sharpness information is localized in low-level features, we argue that this multi-layer architecture is not necessary for FQA. On the contrary, we show that the proposed FocusLiteNN model with a single convolution layer is sufficient to achieve a high accuracy that is on par with the state-of-the-art deep CNN models. Besides performance, the major advantage of using a single-layer CNN is its super low computational complexity. While most CNNs are computationally expensive and do not meet the requirements of high-throughput WSI scanning, FocusLiteNN uses only a fraction of the computational resources.

Once we assume that the sharpness level is uniform within a small patch $\mathbf{X} \in \mathbb{R}^{H\times W\times K}$ of a WSI scan, the sharpness of that patch can be represented by a scalar $y \in \mathbb{R}$. As discussed earlier, we set $H = W = 235$. Our objective is to predict y for each patch in a WSI and summarize the results. Similar to knowledge-based and CNN-based FQA

methods, the process begins with convolving the patch with a set of kernels $\mathbf{\Phi} \in \mathbb{R}^{h \times w \times K \times N}$ to extract the features. Then a non-linear regression function f is applied to deduce the features to a sharpness score. This process is formulated as

$$y = f\left(\sum_{k=1}^{K} \mathbf{\Phi_k} * \mathbf{X_k} + \mathbf{b}\right) \tag{3.4}$$

where, $\mathbf{\Phi}_k \in \mathbb{R}^{h \times w \times N}$ is the convolution kernel for kth input channel. Here, $\mathbf{X_k} \in \mathbb{R}^{H \times W}$ is kth channel of input patch and $\mathbf{b} \in \mathbb{R}^N$ is a bias vector. When doing the convolution, $\mathbf{X_k}$ is repeated N times, which has the shape of $H \times W \times N$. Similarly, $\mathbf{b}$ is repeated $H \times W$ times, which has the shape of $H \times W \times N$. $y \in \mathbb{R}$ is the predicted score of $\mathbf{X}$. As mentioned earlier, color information is crucial in this task since it provides more cues in contrast. Therefore, we set $K = 3$ through this thesis, except for visualization where grayscale input is used ($K = 1$). The 2D convolution operator $*$ is applied with a stride of 5. We set the kernel size to $h = w = 7$ for all experiments. The convolution stride is set to 5 to balance the computational complexity and performance. We refer to the model in (3.4) as N-kernel mode of FocusLiteNN.

Regression Function In knowledge-based FQA methods, both handcrafted and machine learning-based regression models are used. Handcrafted regression models use a predefined function, such as summation or product, to merge the features into scores. The specific form and parameters are either chosen empirically or tuned based on a small set of labeled data. In contrast, machine learning regression models search for the optimal function in a larger functional space through training on a larger set of data. Commonly used machine learning regression/classification models include SVR [90, 91, 92, 111, 118], SVM [35, 34], linear regression [38], logistic regression [52], AdaBoost [15, 51], and decision tree [34, 52], etc. However, these models are usually suboptimal since the feature extraction module and the regression module are not optimized jointly. Some methods are also computationally expensive. For example, the time complexities of SVR and SVM are more than quadratic with the number of samples. This makes the training and inference difficult on large datasets.

In data-driven FQA methods where CNNs are used to extract features and predict

the final scores in an end-to-end fashion, the most commonly used regression model is MLP. MLP consists of several fully connected layers with nonlinear activation functions in between. From a neural network's perspective, adding non-linearity to the model greatly enhances the approximation capability. However, MLPs can have a high number of parameters, especially when dealing with high-dimensional inputs like images. This is because every neuron in one layer is connected to every neuron in the next layer. Consequently, MLPs require more computational resources for training and inference. Furthermore, MLPs are more susceptible to overfitting, especially when the amount of training data is limited.

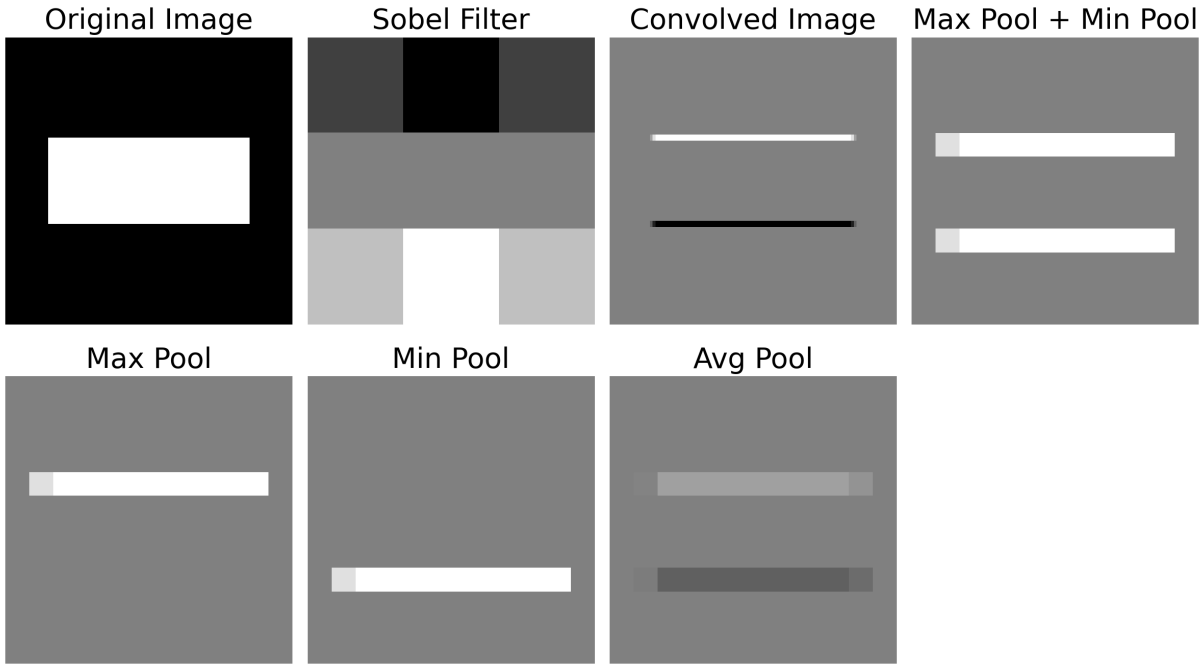


Figure 3.9: How different pooling strategies affect the feature map of a Sobel kernel filtered edge image. While max pooling or min pooling alone can only capture one of the two edges. The max pooling + min pooling strategy captures both the increasing and decreasing edges in the original image. Average pooling produces a blurred feature map where the activation is less significant. In the figures, the value of the minimum pooling is inverted.

To reduce the computational cost, we aim for a minimum design that can still maintain

competitive performance compared with the state-of-the-art CNN models. We defined the regression function as

$$f(\mathbf{x}) = \langle \mathbf{w_1}, \max(\mathbf{x}) \rangle + \langle \mathbf{w_2}, \min(\mathbf{x}) \rangle \tag{3.5}$$

where, $\mathbf{x} \in \mathbb{R}^{\frac{H-h+7}{5} \times \frac{W-w+7}{5} \times N}$ are the feature maps produced by the convolution operation. $\max(\cdot) : \mathbb{R}^{\frac{H-h+7}{5} \times \frac{W-w+7}{5} \times N} \rightarrow \mathbb{R}^N$ and $\min(\cdot) : \mathbb{R}^{\frac{H-h+7}{5} \times \frac{W-w+7}{5} \times N} \rightarrow \mathbb{R}^N$ are spatial-wise maximum and minimum pooling. $\mathbf{w_1}, \mathbf{w_2} \in \mathbb{R}^N$ are trainable parameters. The use of spatial-wise 2D maximum pooling and minimum pooling has several benefits. Firstly, it makes the model capable of capturing extreme kernel responses. Since these operations retain the highest/lowest value in each pooling window, it ensures that the most significant signals are preserved. This reduces redundant information and reduces the noise, which helps the model to focus on the most important aspects of the input. Secondly, it provides a certain degree of translation invariance. By taking the maximum value over a pooling window, the exact location of the extreme activations becomes less important. This is beneficial because the network becomes more robust to slight variations and shifts in the position of features in the input image. Thirdly, it helps in reducing the dimensions of the feature maps, which reduces the computational load for the network. We demonstrate the effectiveness of Eq 3.5 in Fig 3.9. The original rectangle image is convolved with a horizontal Sobel filter. The Sobel filter detects two edges in the image, one increasing and one decreasing. We process this feature map with different pooling strategies: max pooling, min pooling, average pooling, and max + min pooling. While max pooling or min pooling alone can only capture one of the two edges, average pooling is able to capture both. However, it produces a blurred feature map where the activation is less significant. Only the max pooling + min pooling strategy captures both the increasing and decreasing edges which stand out from the background.

Loss Function

The loss functions can be generally categorized into two classes: for classification and for regression. Most classification tasks use cross entropy as the loss function, where the categories are assumed to be nominal. Nominal classes are unordered and mutually exclusive. However, in the case of FQA, categories representing different focus levels are ordered. We refer to this kind of variable as ordinal. For example, some datasets categorize blur levels into five categories: very poor, poor, ok, fairly good, very good. In this scenario,

ordinal classification [67] might be handy since these categories are ordered but the distance between classes is hard to define. This rank information provides more cues for the network and generally results in better performance [66].

However, for z-stack-based FQA datasets, the distance between any blur levels is clearly defined. In this scenario, regression losses will be a better choice because it provides information about the error distance. The major difference between ordinal classification and regression is that the distances between the classes are unknown for the ordinal case. For regression, the most commonly used loss function is MAE and MSE. The difference between them is that MSE is smooth around 0, but it is more sensitive to larger differences, such as outliers. MAE is more robust to outliers, but it is less smooth around 0. To take advantage of both loss functions, Huber loss and smooth L1 loss use MSE in small value region while using MAE in large value region. Huber loss is defined as

$$Huber(x_i, y_i) = \begin{cases} 0.5\,(x_i - y_i)^2\,, & \text{if } |x_i - y_i| < \delta \\ \delta\,(|x_i - y_i| - 0.5\delta)\,, & \text{otherwise} \end{cases} \tag{3.6}$$

where δ specifies the threshold at which to change switch from MAE loss to MSE loss.

More recently, PLCC is also utilized as a loss function in regression problems [174]. PLCC measures the linear correlation between two sets of data, which is a popular metric for evaluating IQA models on subjectively rated datasets. It is formulated as

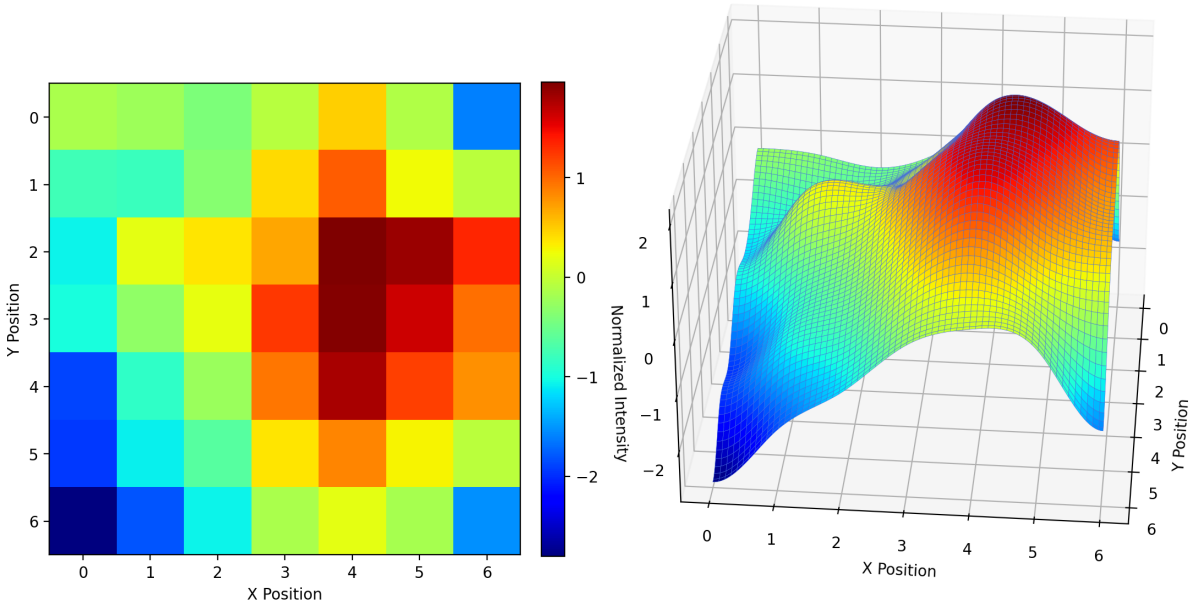
$$PLCC(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.7}$$

where $\mathbf{x} = \{x_i | i = 1, 2, ..., n\}$ and $\mathbf{y} = \{y_i | i = 1, 2, ..., n\}$ are the predicted scores and ground truth scores, respectively. $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ and $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ are the mean of $\mathbf{x}$ and $\mathbf{y}$, respectively. When used as a loss function, we normally maximize PLCC or minimize the negative PLCC. Compared to error-based loss functions such as MSE, correlation-based loss functions such as PLCC have several advantages [175]. Firstly, PLCC makes the model transform features in such a way that the linear correlation w.r.t to target is maximized. Secondly, the ranges of the prediction and ground truth data do not have to be aligned. In

certain models, the range of the prediction can be bounded and does not match that of the ground truth. In this scenario, error-based loss functions will fail to provide meaningful gradients. Thirdly, PLCC is regularized to the range from -1 to 1, which decreases the influence of outliers and avoids the exploding gradient problem. Fourthly, PLCC is one of the major evaluation metrics used in FQA, ISA and IQA. This alignment ensures that the optimization directly contributes to improving the metric of interest.

We evaluated various loss functions under both classification and regression scenarios. The results are shown in Sec 3.4. Since the training dataset for FocusLiteNN is z-stack-based, we finally choose the negative PLCC as the loss function for its overall performance.

Filter Visualization



(a) 2D spatial representation

(b) 3D spatial representation

Figure 3.10: FocusLiteNN (1-kernel grayscale) filter visualization: (a) 2D spatial representation, (b) 3D spatial representation

To better understand the FocusLiteNN model, we visualize the $\mathbf{\Phi} \in \mathbb{R}^{7\times7\times N}$ of the trained 1-kernel model ($N = 1$) in both spatial and frequency domains. Due to the space

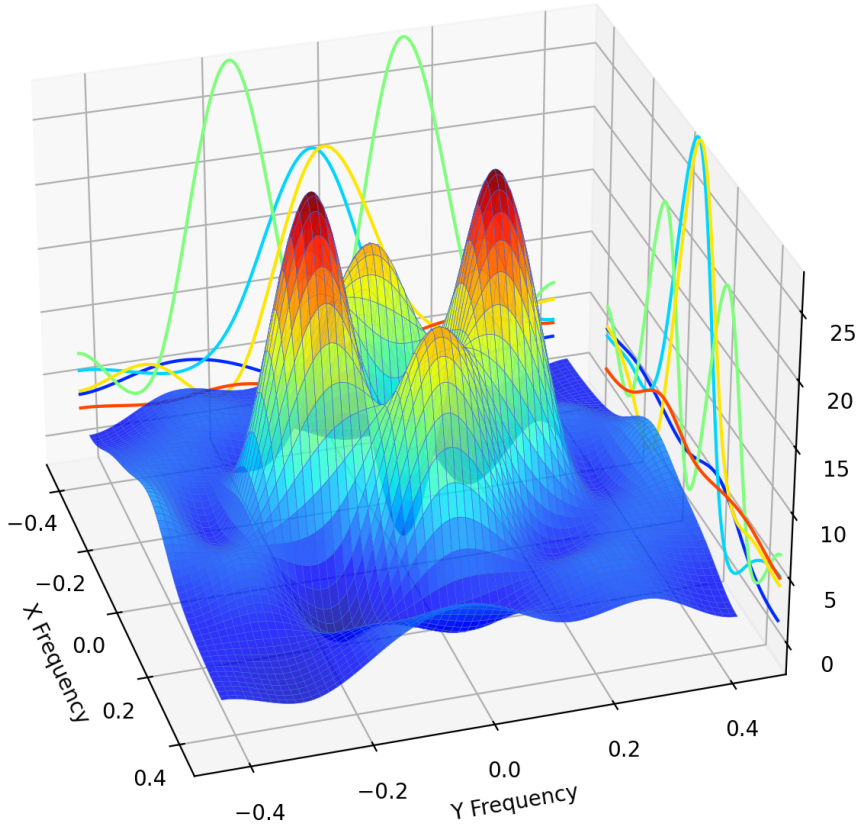


Figure 3.11: FocusLiteNN (1-kernel grayscale) filter visualization: shifted FFT amplitude.

constraint, we set the input channel $k = 1$ for the purpose of visualization, which means grayscale images are used to train this demo model. In all other experiments, we use color images where $k = 3$. In Fig 3.10, we illustrate the spatial representation of this kernel in both 2D (a) and 3D (b). For comparison, we also illustrate the 3D spatial representation of the vertical Sobel filter and LoG filter in Fig 3.14 (a) and (b), respectively. Since all three filters have positive and negative entries, they are capable of capturing pixel changes, such as edges and finer textures, to some extent. The Sobel filter is known to approximate

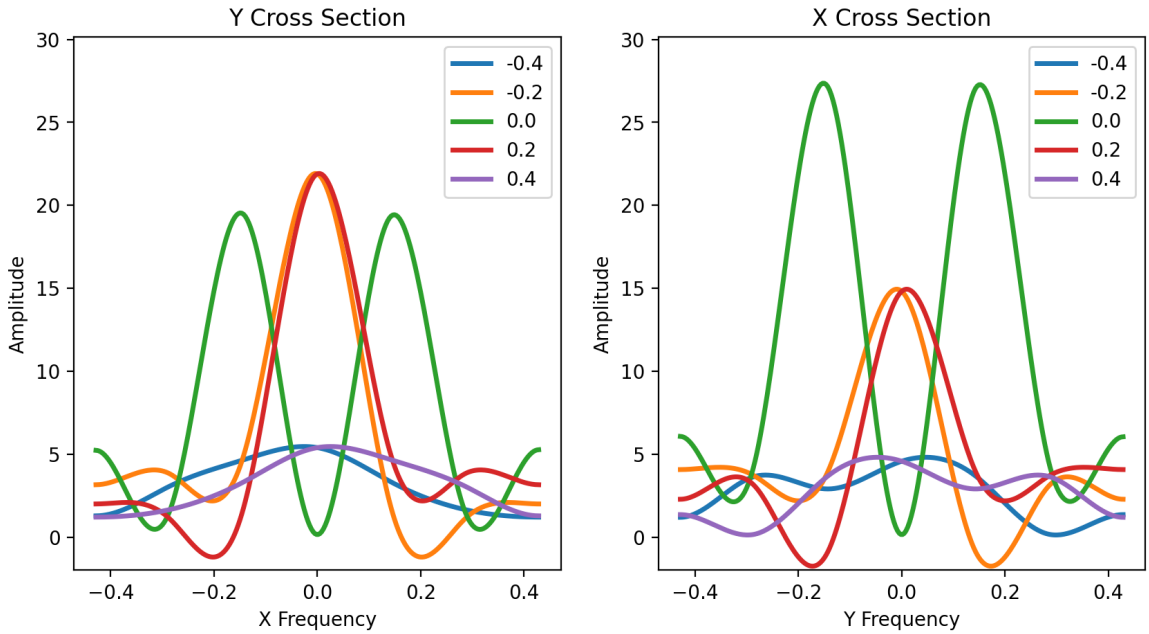


Figure 3.12: FocusLiteNN (1-kernel grayscale) filter visualization: vertical and horizontal cross sections for FFT amplitudes.

the image gradient (first-order derivative), which is used to detect edges. The LoG filter approximates the second-order derivative, which is used to detect edges and finer textures. The vertical Sobel filter has a clear directional component in its spatial representation. The LoG filter has a symmetrical bowl-like spatial representation, showing its isotropic nature. The FocusLiteNN kernel's spatial representation includes a combination of isotropic and anisotropic characteristics, which is tailored for FQA.

Analyzing filters in the frequency domain also provides a lot of insights. The 3D FFT amplitude of the FocusLiteNN kernel is shown in Fig 3.11, and its vertical and horizontal cross sections are shown in Fig 3.12. For comparison, the 3D FFT amplitude of the vertical Sobel filter and LoG filter are also shown in Fig 3.14 (c) and (d), respectively. It is clear that both Sobel and LoG filters, like the FocusLiteNN kernel, emphasize certain frequencies over others. This is evident in the peaks of their FFT amplitude representations. The FocusLiteNN kernel seems like a bandpass filter that has both directional and omnidirec-

tional responses. The response pattern is more complex and balances the trade-off between detecting fine details (like the LoG) and preserving important structural information (like the Sobel). Unlike the Sobel and LoG filters where the accentuating frequencies are hand-crafted, the bandpass characteristics of the FocusLiteNN kernel are learned from the WSI data, making the frequency selection tailored for this task.

We further show in Fig 3.13 the feature maps produced by convolving the image with the learned FocusLiteNN kernel and the Sobel filter. In both feature maps, it is clear that they capture edge-related information. In in-focus regions, we can find more extreme activations shown as blue and red. Whereas in out-of-focus regions, most activations are less prominent, shown as green. In the feature map created by the FocusLiteNN learned filter, it is easy to find that it captures edges better than the Sobel filter.

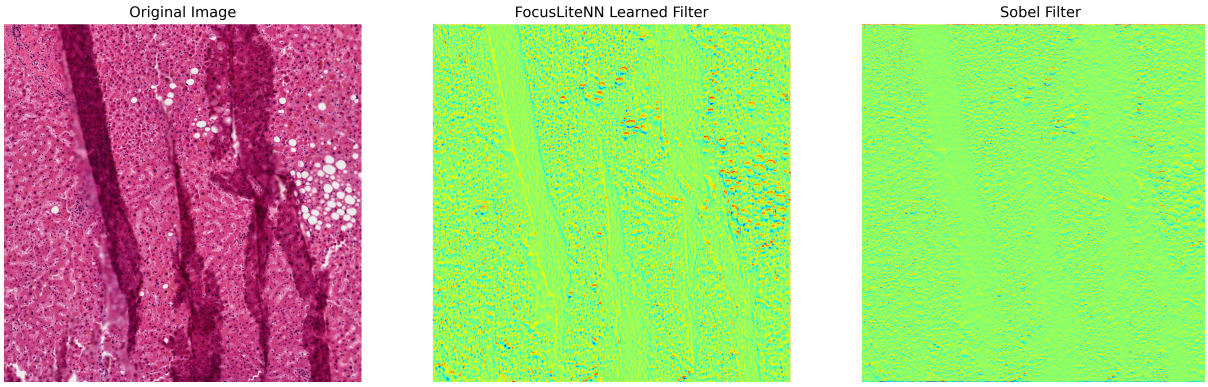


Figure 3.13: The feature maps produced by convolving the image with the learned FocusLiteNN kernel and the Sobel filter.

3.3 TCGA@Focus Dataset

The development of data-driven FQA in digital pathology heavily relies on the selection of datasets for training. While the CNN models perform very well on the training dataset, the ultimate question is how well the models can be transferred to other datasets for evaluation. This is of paramount importance in digital pathology where the models should be capable

Dataset	Year	Type	Mag	# Scanners	# Slides	# Organs	# Stains
DeepFocus [60]	2018	Z-Stack	40×	1	16	N.A.	4
FocusPath [2]	2019	Z-Stack	40×	1	9	9	8
TCGA@Focus [176]	2020	Manual	40×	> 1	1000	52	N.A.

Dataset	# Patches	# IF	# OOF	Patch Size	# Equivalent Patches
DeepFocus	$204,000$	$108,000$	$96,000$	64×64	$12,750$
FocusPath	$8,640$	$1,620$	$7,020$	1024×1024	$8,640$
TCGA@Focus	$14,371$	$11,328$	$3,043$	1024×1024	$14,371$

Table 3.1: The details of the proposed TCGA@Focus dataset, compared with the only two public available FQA datasets. IF and OOF are the abbreviations for in-focus and out-of-focus, respectively. The number of equivalent patches is calculated as the number of 1024×1024 patches.

of (a) accurately predicting focus scores on the slides regardless of tissue structures and staining protocols; and (b) accounting for color disparities that could be caused by WSI scanner variations and tissue preparation in different pathology labs.

However, as reviewed in Sec 2.1.2, only two datasets are publicly available. Each of the two datasets is captured using only one scanner in one lab. As detailed in Table 3.1, DeepFocus [60] collected four slides with different stains for each of the four patients, resulting in a total of 16 slides. FocusPath [2] collected one slide for each of the nine organs and eight types of stains were used. The number of slides and the diversity of organs in the two datasets are limited. Both datasets are limited to slides prepared by one lab and scanned using one WSI scanner. Since the pathological tissue preparation protocol and imaging platform can affect the quality and appearance of the slide, the diversity of blur distortion and stains is also limited. In addition, DeepFocus [60] is constrained by a low resolution of 64×64 pixels, which may not capture the field of view required for accurate analysis. FocusPath offers higher resolution patches at 1024×1024 pixels, but its dataset is considerably smaller with only $8,640$ patches.

To address the above-mentioned limitations and provide a more comprehensive resource for FQA, we proposed a new manually labeled FQA dataset, TCGA@Focus [3], consisting of 1000 WSI of 52 organ types selected from the TCGA repository in SVS format. TCGA is a project to catalog the genetic mutations responsible for cancer using genome sequencing and bioinformatics [10]. It is co-managed by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI), which are both part of the National Institutes of Health (NIH). While TCGA's efforts have primarily focused on the genomics data, it also hosts over $30,000$ WSIs spanning 33 cancer types along with associated pathology reports, clinical data, surgery and radiation treatment information, and genomic information [75]. Diversity is the highlight of this dataset. The data are collected from patients of different races, ethnicities, ages and genders. The tissues are prepared using different preparation protocols and scanned using different equipment in labs across the United States. The statistical distribution of the number of slides per organ site is shown in Figure 3.15. The distribution of the TCGA@Focus dataset w.r.t. the organ type is relatively uniform. Note that the diversity of the organ types here is important to include a wide spectrum of tissue textures and color information caused by variations in staining and WSI scans.

Since these slides are prepared for the purpose of clinical diagnosis instead of FQA, they contain more diverse physical artifacts, such as tissue fold, bubbles, markers, etc, which appear as different kinds of out-of-focus blur in the final images. In addition, the distortions are often nonuniformly distributed across the spatial domain: most parts stay in-focus while few areas suffer from out-of-focus blur. On the other hand, the out-of-focus blur in z-stack-based datasets is mainly caused by setting an incorrect focus plane to the specimen. Also, the distortion is often spatially invariant. In conclusion, while the z-stack method makes the labeling process more efficient, the distortions captured are different from real-world applications in both style and distribution.

With the $1,000$ slides collected and pre-screened, the next step is to extract patches from the slides and label them. Since most regions within the slides are free from distortion, extracting patches uniformly from the slides will cause an imbalance in the distribution of labels. To construct the dataset, we inspect each slide and select in-focus and out-of-focus

[3]The data is available at https://zenodo.org/record/3910757

regions-of-interests (ROI). Regions with a high portion of background are discarded during the process. Then patches are randomly extracted from these regions. To ensure a large field of view, high-resolution patches at 1024×1024 pixels are extracted from the original WSIs, allowing for finer detail and more accurate quality assessment. The patch examples are shown in Figure 3.16. Then the patches are classified manually by experts into two different categories: *"in-focus"* and *"out-focus"*, corresponding to the binary ground truth scores of "1" and "0", respectively. The compiled dataset contains $14,371$ image patches in total, where $11,328$ patches are labeled in-focus and $3,043$ patches out-focus. A label indicating whether the out-of-focus is caused by markers is also provided. The manual annotation process ensures the high quality of data, which is vital for training or testing reliable and accurate models for medical image analysis.

3.4 Experiments

3.4.1 Implementation Details

As described in Sec 2.1.2, only two datasets are publicly available. We choose the FocusPath dataset [2] as the training dataset for its larger patch size. The FocusPath[4] dataset [2] contains $8,640$ patches of 1024×1024 image extracted from nine different stained slides. The WSIs are scanned by Huron TissueScope LE1.2 [77] using 40X optics lens @0.25μm/pixel resolution. For each location, the autofocus system first determines the focus distance with the sharpest image. Then the stage moves the tissue both away from and toward the objective lens in incremental steps to create several out-of-focus patches. Each patch is associated with a z-level (referred to as slice in the filename) ranging from 1 to 16 with the sharpest level at 8 or 9. 1 and 16 are the most blurry levels. Some examples are shown in Figure 2.4 for the same patch at different z-levels.

However, the z-level can not be used as the ground truth for supervised training since it does not correlate linearly with the level of blur. To create the ground truth, we need to map the z-level to the blur level. Since the z-level is linearly correlated with the focus

[4]The data is available at https://zenodo.org/record/3926181

distance and the level of blur can be characterized by the radius of the circle of confusion, what we need to study is the correlation between the focus distance and the circle of confusion. In optics, the Airy disk is the in-focus image of a point light source passing through a circular aperture, which is the 2D projection of the PSF on the image plane. Due to diffraction, the Airy disk is a bright circular disk surrounded by an alternating series of bright and dark rings. On the other hand, the circle of confusion is a simplified version of the Airy disk where diffraction is ignored. When in focus, the circle of confusion reduces to a single point on the image plane. When out-of-focus, it becomes a bright circular disk. An illustration of the circle of confusion is shown in Figure 3.17.

The radius of the circle of confusion is given by

$$\sigma = A\frac{|F-d|f}{2(F-f)d} \tag{3.8}$$

where d is the focusing distance, F is the objective distance, f is the focal length, A is the diameter of the aperture. Take Nikon's CFI Plan Fluor 40X objective lens [5] for instance, it numerical aperture is $NA = 0.75$. Since the diameter of the aperture A and the focal length f are not provided by the manufacturer, we estimate them as $A = 11mm$ and $f = 5mm$. Based on these parameters, we plot the radius of the circle of confusion as a function of the focusing distance d in Fig 3.18. If we ignore the diffraction and only consider the geometric optics, the radius equals zero when the focus distance matches the designed in-focus distance, shown as the red dotted line. It is easy to find that the circle of confusion becomes bigger when the offset of the focus distance w.r.t to the in-focus distance becomes larger. It is clear that the curve is almost symmetric about the in-focus position. This means that the blur level scales linearly at the same speed in both directions. According to this, we convert the z-level to the blur level y using the following equation

$$y = |z - z^*| \tag{3.9}$$

where $z \in \{1, 2, \cdots 16\}$ is the z-level of the patch, and $z^* \in \{1, 2, \cdots 16\}$ is the in-focus z-level at this location that creates the sharpest image. z^* is usually determined by the

[5] https://www.microscope.healthcare.nikon.com/products/optics/selector/comparison/-1828

autofocus system or by applying a sharpness assessment algorithm on the z-stack. For example, in the FocusPath dataset [2], the sharpest z-level was determined by the HVS-MaxPol metric.

Since the FocusPath includes diverse color stains compiled with different tissue structures, this makes the dataset well suited for the development of data-driven FQA models. Furthermore, we hypothesize that the diversity of color stains greatly helps generalize the CNN training to different tissue structures and color spectrum–no color augmentation is required such as in [177].

We adopt five different categories in knowledge-based methods for the experiments using (1) human visual system: Synthetic-MaxPol [58], and HVS-MaxPol-1/HVS-MaxPol-2 [2], (2) microscopy lens modeling: FQPath [57], (3) natural image statistics: MLV [93], SPARISH [116]; and (4) signal processing based: GPC [108] and LPC [107]. For data-driven methods we select a diverse range of CNN models in terms of architecture complexity using EONSS [174] with four conv layers developed for the purpose of Image Quality Assessment (IQA), as well as DenseNet-13 [76] (eight conv layers) and variations of ResNet [64] (8, 48, and 99 conv layers) developed for computer vision applications. We evaluate selected FQA models in terms of statistical correlation and classification performance as well as computational complexity on the FocusPath and TCGA@Focus datasets. At the end, we also show the heat maps generated by these models on a sample image.

All CNNs are re-trained on the FocusPath dataset with the same pre-processing techniques, optimizer and loss function. The FocusPath dataset is randomly split into a train (60%) - validation (20%) - test (20%). The validation subset is used to determine the hyper-parameters. Training and testing are repeated in 10 folds of splits and the average performance is reported. All models are transferred to TCGA@Focus dataset for evaluation. The input dimensions for all CNNs are set to $235 \times 235 \times 3$. During testing, we densely sample the original patches with a stride of 128×128 and the average score is taken as the overall sharpness. Adam optimizer is utilized for all models. For FocusLiteNN, the learning rate is set to 0.01 with decay interval of 60 epochs. For other models, the learning rate is set to 0.001 with decay interval of 40 epochs. Each model is trained for 120 epochs to ensure convergence. The Pearson Linear Correlation Coefficient (PLCC) is used as the loss function for all models. PLCC bounds the loss value between -1 and 1, which helps to

stabilize the training process.

3.4.2 Performance Evaluation

The metrics used to evaluate the performances are Spearman's Rank Correlation Coefficient (SRCC), PLCC, Area Under the Curve of the Receiver Operating Characteristic curve (ROC), Area Under the Curve of the Precision Recall Curve (PR). SRCC measures the monotonicity between the predicted sharpness score and the absolute z-level, while PLCC measures the linear correlation between them. When measuring ROC and PR on the FocusPath dataset, we first binarize the z-levels by considering all patches with absolute z-level 0, 1, 2 as sharp and those equal or larger than 2 as blurry. The results are shown in Table 3.2.

Type	Model	FocusPath				TCGA@Focus		Size	Time (sec)	MMACs
		SRCC	PLCC	ROC	PR	ROC	PR			
Data-driven based	FocusLiteNN (1-kernel)	0.8766	0.8668	0.9468	0.9768	0.9310	0.8459	**148**	**0.017**	**0.3**
	FocusLiteNN (2-kernel)	0.8782	0.8686	0.9481	0.9770	**0.9337**	0.8499	**299**	**0.019**	**0.7**
	FocusLiteNN (10-kernel)	0.8931	0.8857	0.9542	0.9802	0.9322	**0.8510**	**1.5K**	**0.019**	**3.3**
	EONSS [174]	0.9009	0.8951	0.9540	0.9799	0.9000	0.8473	123K	0.063	13.7
	DenseNet-13 [76]	**0.9253**	**0.9197**	**0.9662**	0.9849	**0.9386**	**0.8646**	193K	0.355	364.4
	ResNet-10 [64]	**0.9278**	**0.9232**	**0.9671**	**0.9853**	0.9292	**0.8559**	4.9M	0.334	1044.7
	ResNet-50 [64]	**0.9286**	**0.9244**	**0.9676**	**0.9855**	**0.9364**	0.8144	24M	1.899	4819.0
	ResNet-101 [64]	0.9242	0.9191	0.9644	**0.9840**	0.9320	0.8447	43M	2.655	9104.0
Knowledge based	FQPath [57]	0.8395	0.8295	0.9375	0.9739	0.7483	0.6274	N.A.	0.269	N.A.
	HVS-MaxPol-l [2]	0.8044	0.8068	0.9400	0.9743	0.7118	0.5622	N.A.	0.257	N.A.
	HVS-MaxPol-2 [2]	0.8418	0.8330	0.9434	0.9757	0.7861	0.6721	N.A.	0.458	N.A.
	Synthetic-MaxPol [58]	0.8243	0.8139	0.9293	0.9707	0.6084	0.4617	N.A.	0.841	N.A.
	LPC [107]	0.8375	0.8321	0.9223	0.9681	0.5576	0.4564	N.A.	7.510	N.A.
	GPC [108]	0.7851	0.7602	0.9095	0.9604	0.4519	0.2830	N.A.	0.599	N.A.
	MLV [93]	0.8623	0.8528	0.9414	0.9758	0.8235	0.6943	N.A.	0.482	N.A.
	SPARISH [116]	0.3225	0.3398	0.7724	0.8875	0.7293	0.6414	N.A.	4.853	N.A.

Table 3.2: SRCC, PLCC, ROC-AUC, PR-AUC Performance of 16 NR-ISA Metrics on FocusPath Dataset and TCGA@Focus Dataset. The number of parameters, average processing time, and computational complexity are also reported.

On the FocusPath dataset, the overall performance of DenseNet-13 [76], ResNet-10 [64], ResNet-50 [64] and ResNet-101 [64] in all 6 metrics are the best and are similar

	FocusLiteNN	EONSS	DenseNet13	ResNet50	FQPath	HVS-MaxPol	Synth-MaxPol	LPC	GPC	MLV	SPARISH
FocusLiteNN	-	0	0	0	1	1	1	1	1	1	1
EONSS	1	-	0	0	1	1	1	1	1	1	1
DenseNet13	1	1	-	-	1	1	1	1	1	1	1
ResNet50	1	1	-	-	1	1	1	1	1	1	1
FQPath	0	0	0	0	-	1	1	-	1	0	1
HVS-MaxPol	0	0	0	0	0	-	-	0	1	0	1
Synth-MaxPol	0	0	0	0	0	-	-	0	1	0	1
LPC	0	0	0	0	-	1	1	-	1	0	1
GPC	0	0	0	0	0	0	0	0	-	0	1
MLV	0	0	0	0	1	1	1	1	1	-	1
SPARISH	0	0	0	0	0	0	0	0	0	0	-

Table 3.3: Statistical significance testing of FQA methods on the FocusPath dataset using prediction residuals. 1 means that the method is statistically better than the method in the column, 0 means that it is statistically worse, and - means that it is statistically indistinguishable.

to each other. Assuming that the testing subset of FocusPath is drawn from the same distribution as the training subset, this observation shows that those data-driven-based models with more parameters can fit the distribution of training data better. ResNet-50, the best performer among deep CNN-based models, outperforms the 10-kernel model, the best performer among shallow CNN-based models, by 3.5% in SRCC and 2% in ROC. To visualize the statistical correlation of all models, the scatter plots of the predicted scores versus z-levels on the FocusPath testing subset are shown in Fig 3.19. We can see that the monotonicity and linearity between the prediction and ground truth are best preserved in deep CNN base models. The statistical significance testing is also performed on the FocusPath dataset. The results are shown in Table 3.3. It can be seen that FocusLiteNN outperforms all knowledge-based methods.

All models are also evaluated on the TCGA@Focus dataset to study the transferability performance where no training is involved. Here, DenseNet-13 [76] achieves the highest scores on both ROC-AUC and PR-AUC. While the overall performance of the deep CNN-based models are still in the top tier, the gap between them and the shallow CNNs are getting smaller compared with the performance difference on the FocusPath dataset: ResNet-50 only outperforms the FocusLiteNN (10-kernel) model by 0.4% in terms of ROC.

Kernel	SRCC	PLCC	KRCC	ROC-AUC	PR-AUC
Learned (ours)	**0.8766**	**0.8668**	**0.7132**	**0.9468**	**0.9768**
LoG	0.8043	0.7957	0.6281	0.9225	0.9641
Laplacian	0.7056	0.7264	0.5272	0.9212	0.9541
Sobel	0.2196	0.1538	0.2100	0.6174	0.7922

Table 3.4: A performance comparison of using fixed kernels in FocusLiteNN.

Regression Model	SRCC	PLCC	KRCC	ROC-AUC	PR-AUC	Time
Max+Min Pooling (ours)	**0.8766**	**0.8668**	**0.7132**	**0.9468**	**0.9768**	**1**
MLP	0.7282	0.7184	0.5459	0.8771	0.9418	20
SVR	0.7984	0.7897	0.6219	0.9164	0.9584	3745
RBFNet	0.6649	0.6967	0.4896	0.8832	0.9399	6619

Table 3.5: A performance comparison of regression models: Max&Min Pooling, MLP, SVR, RBFNet. The features used are the same for all regression models. The (inference) time is calculated relative to the Max&Min Pooling model. The feature extraction time is excluded from this test. ROC and PR are calculated based on binary classification.

Distribution of the predicted scores on the TCGA@Focus dataset and their ground truth labels, as well as the classification thresholds for all models, are also shown in Fig 3.20.

3.4.3 Ablation Study

3.4.4 Computational Complexity Analysis

The testing image is $1024 \times 1024 \times 3$ 8-bit in the FocusPath dataset. Two experiments are conducted, the first one is ROC-AUC on the TCGA@Focus dataset versus CPU time (Fig 3.23 left). To fairly compare the computational complexity, all models are running on an Intel i9-7920X @ 2.90GHz with 32 GB memory. Image reading time is excluded from the CPU time, but the pre-processing time for each model, such as dense sampling, is measured. The MontCarlo simulation is done for 100 times and the average is reported. The second experiment is ROC-AUC on the TCGA@Focus dataset versus number of model

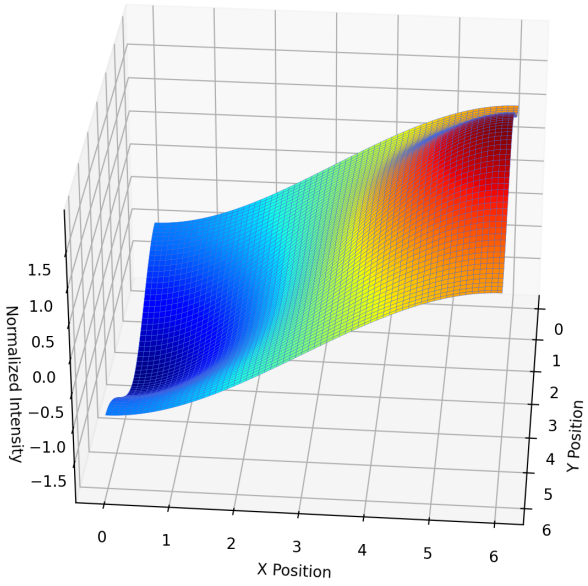
Loss	SRCC	PLCC	KRCC	ROC-AUC	PR-AUC	Maintain Range
PLCC	**0.8766**	**0.8668**	**0.7132**	**0.9468**	**0.9768**	No
CCC	0.8677	0.8025	0.7015	0.9432	0.9759	Yes
MSE	0.8680	0.8576	0.7012	0.9440	0.9763	Yes
MAE	0.8622	0.8511	0.6938	0.9426	0.9756	Yes

Table 3.6: A performance comparison of loss functions in training the FocusLiteNN (1-kernel) model: PLCC, CCC, MSE, MAE, multi-class Cross Entropy and multi-class Ordinal Cross Entropy. ROC and PR are calculated based on binary classification.

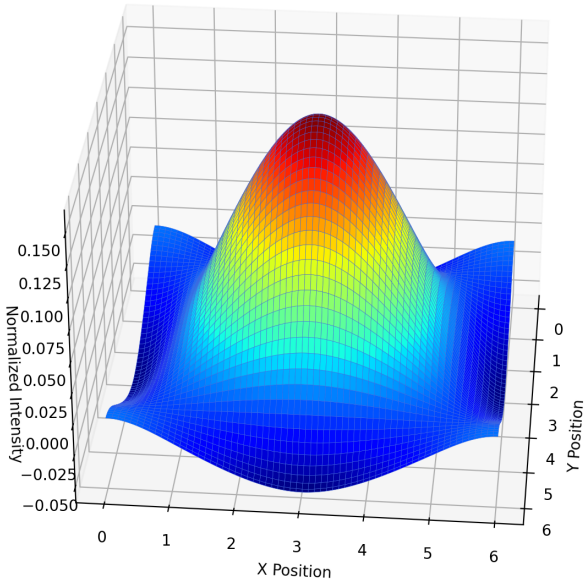
parameters (Fig 3.23 right). We count the number of trainable parameters of the data-driven models and plot the numbers against their performance. We can clearly see that the 1-kernel model outperforms others by a large margin in terms of both CPU time and model size: it outperforms the second fast model EONSS [174] by 3.4% in terms of ROC-AUC, but consuming only 27% of its CPU time with 0.1% of its model size.

3.4.5 Heat Map Visualization

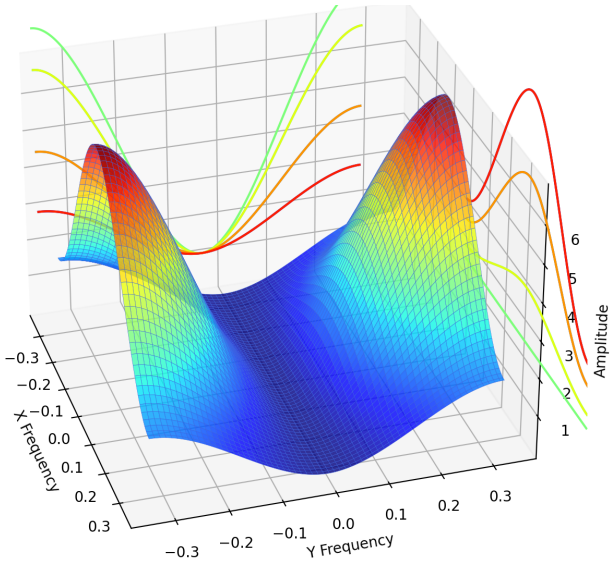
To better visualize the model outputs, we generate heat maps for each model, as shown in Fig 3.22. For all models, we densely sample 235×235 patches from the WSI scan with a stride of 128×128 for scoring and interpolated accordingly. These scores are then mapped to colors and overlaid on the grayscale version of the scan. The most blurry parts are in the upper left corner, lower right corner, and in the circle in the middle. The vertical strip taken up $\frac{1}{3}$ of the space is in focus. In Fig 3.22, we showed the relative blurriness level within a scan by normalizing the scores to the range 0 to 1 before color mapping. Knowledge based models and FocusLiteNN prefer to predict the entire scan as more blurry even for in focus part. Deep CNN-based models such as EONSS [174], ResNet [64] and DenseNet [76] are less aggressive and can identify in focus regions, which are more perceptually accurate.



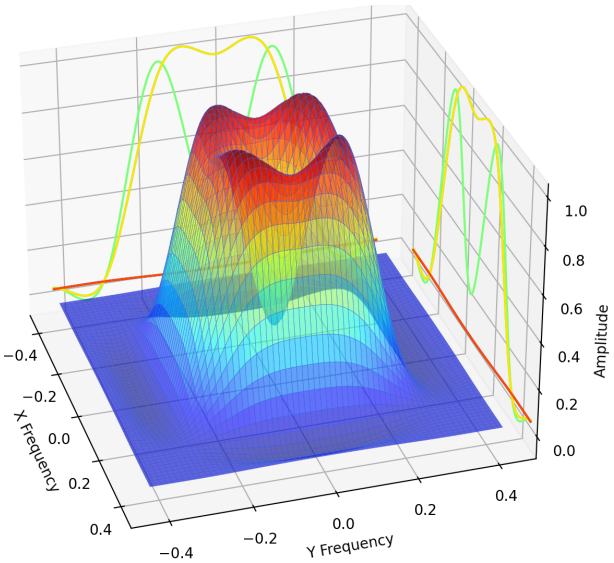
(a) 3D spatial representation of Sobel



(b) 3D spatial representation of LoG



(c) FFT amplitude of Sobel



(d) FFT amplitude of LoG

Figure 3.14: (a) 3D spatial representation of the vertical Sobel filter, (b) 3D spatial representation of the LoG filter (c) 3D FFT amplitude of the vertical Sobel filter, (d) 3D FFT amplitude of the LoG filter

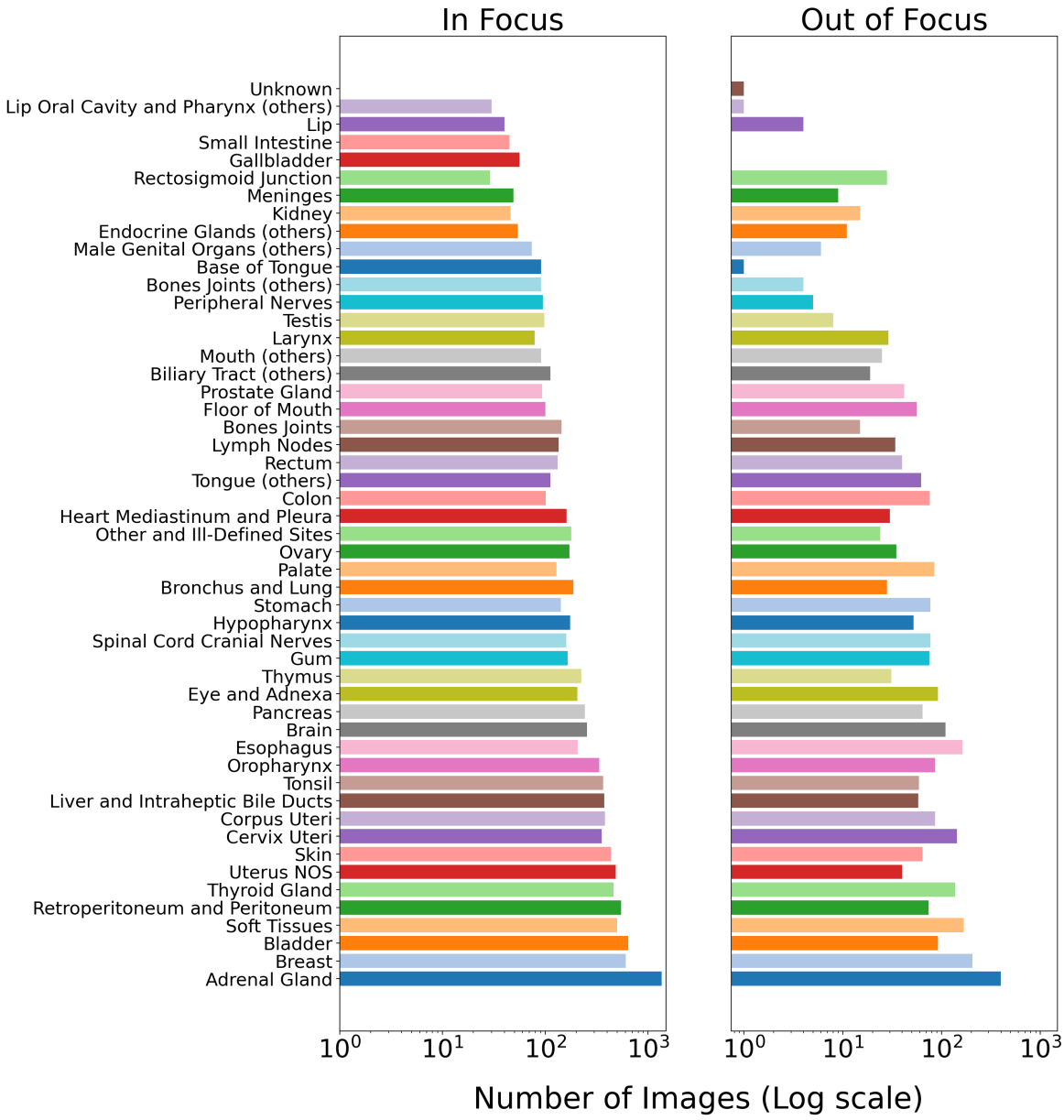


Figure 3.15: Organ distribution of in-focus and out-of-focus images of the TCGA@Focus dataset.

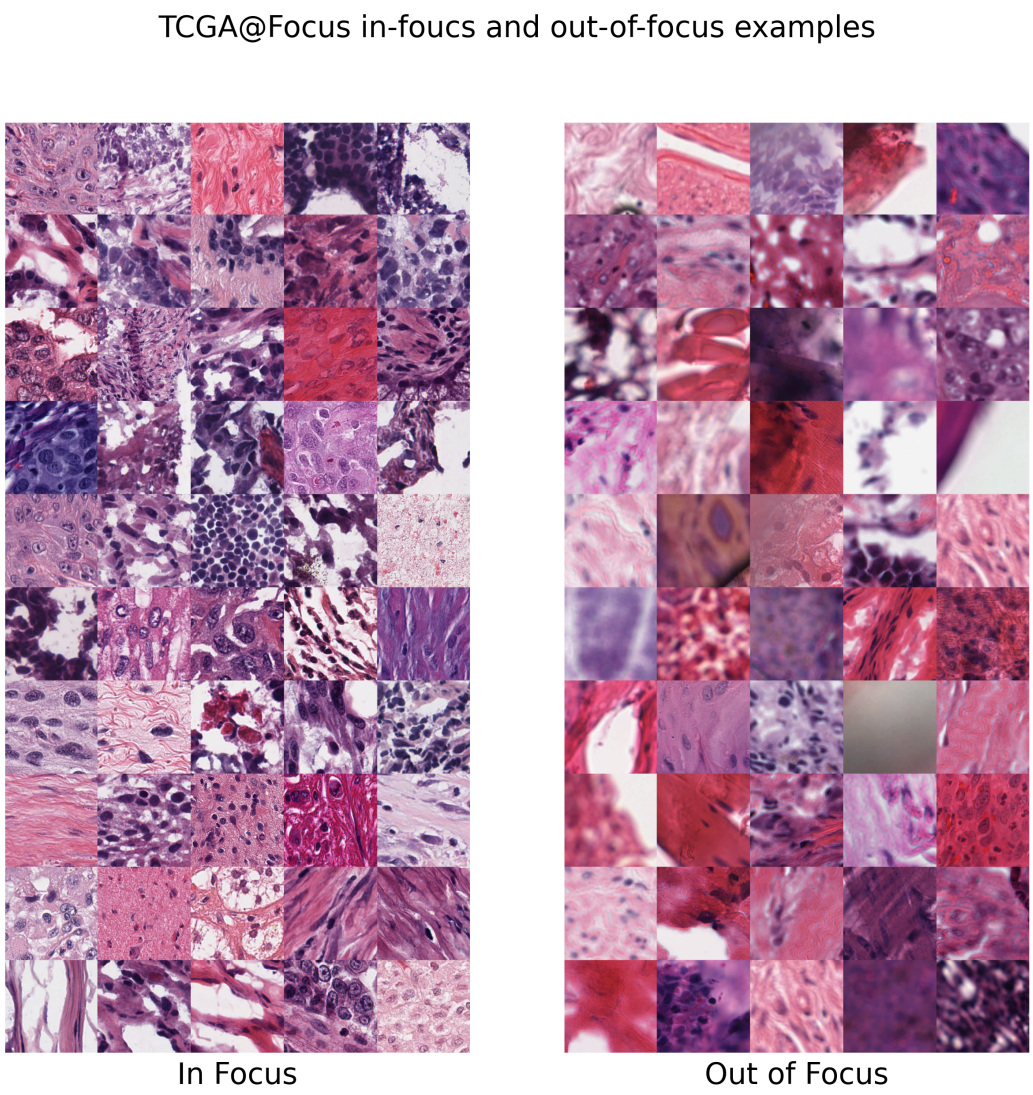


Figure 3.16: In-focus and out-focus examples of the TCGA@Focus dataset.

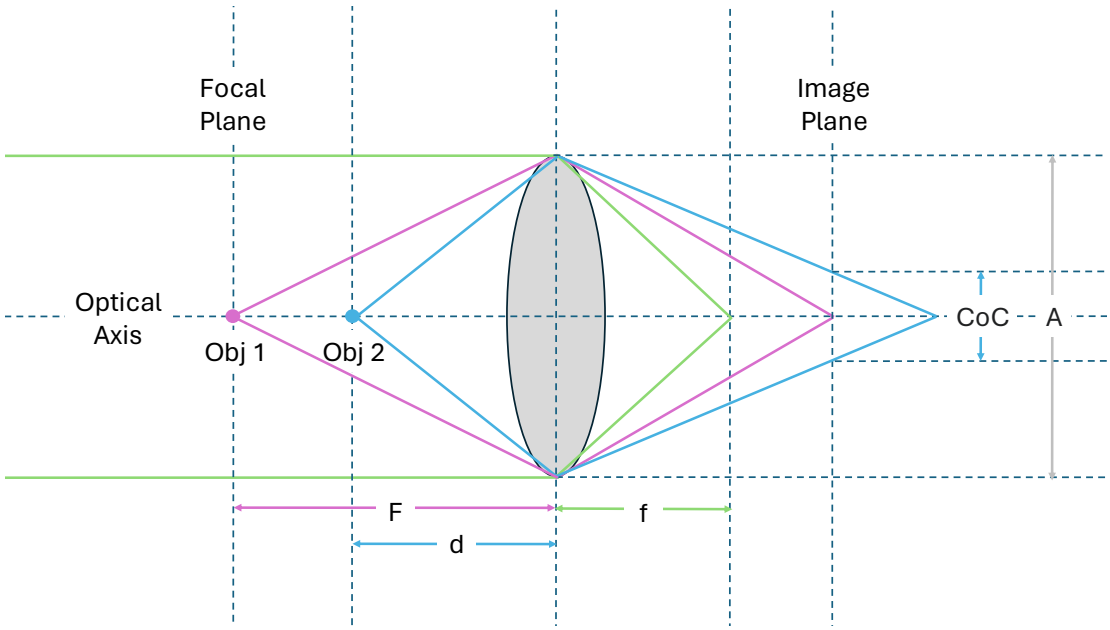


Figure 3.17: The illustration of the circle of confusion when the focus distance does not match the in-focus distance.

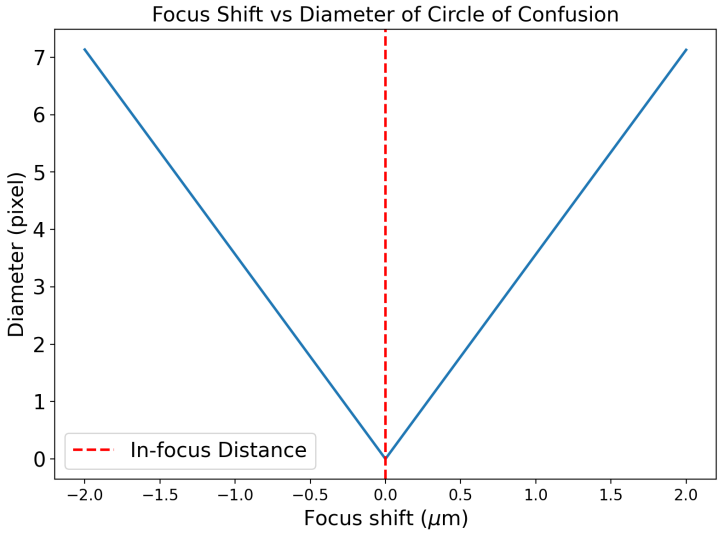
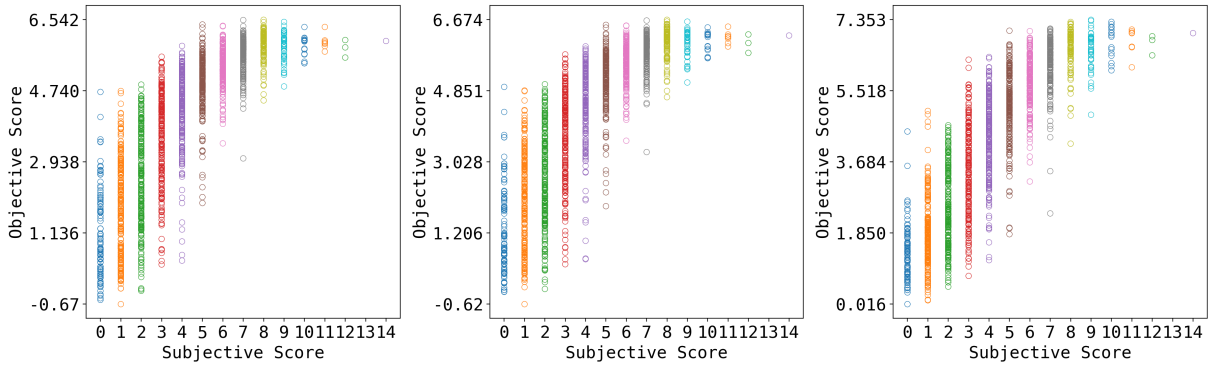


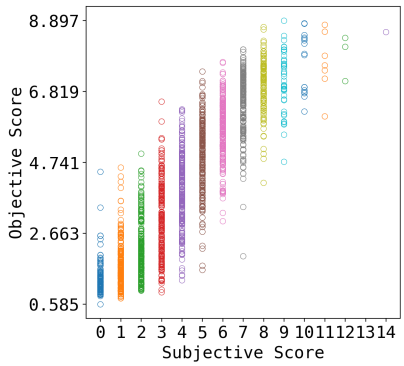
Figure 3.18: Focus distance vs radius of the circle of confusion. It is easy to find that the radius is almost symmetric around the in-focus distance.



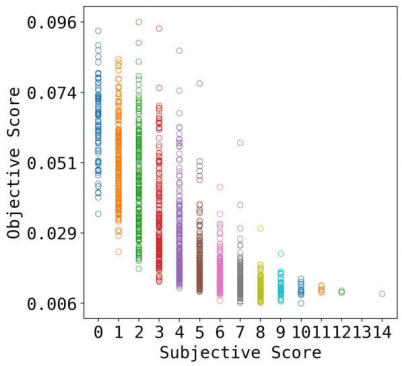
(a) FocusLiteNN (1-kernel)

(b) FocusLiteNN (2-kernel)

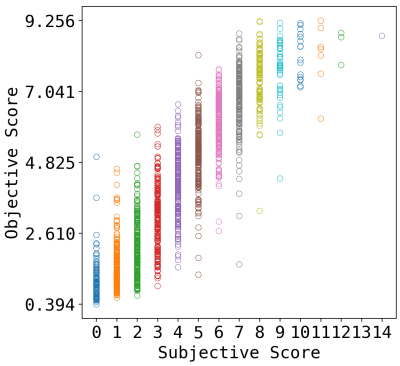
(c) FocusLiteNN (10-kernel)



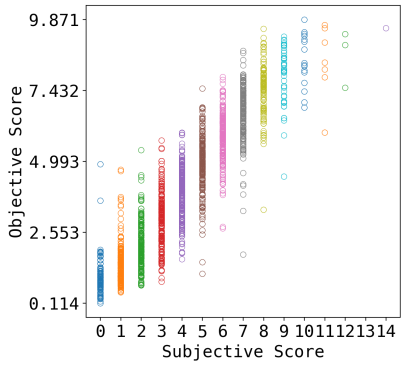
(d) EONSS [174]



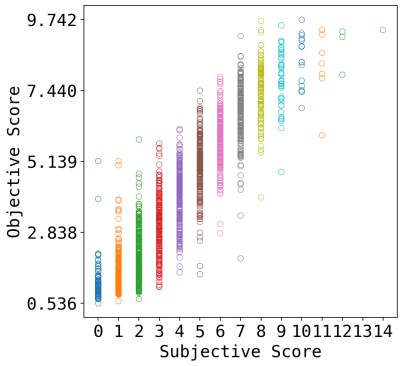
(e) MLV [93]



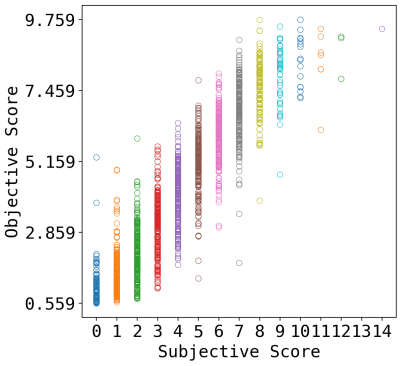
(f) DenseNet-13 [76]



(g) ResNet-10 [64]

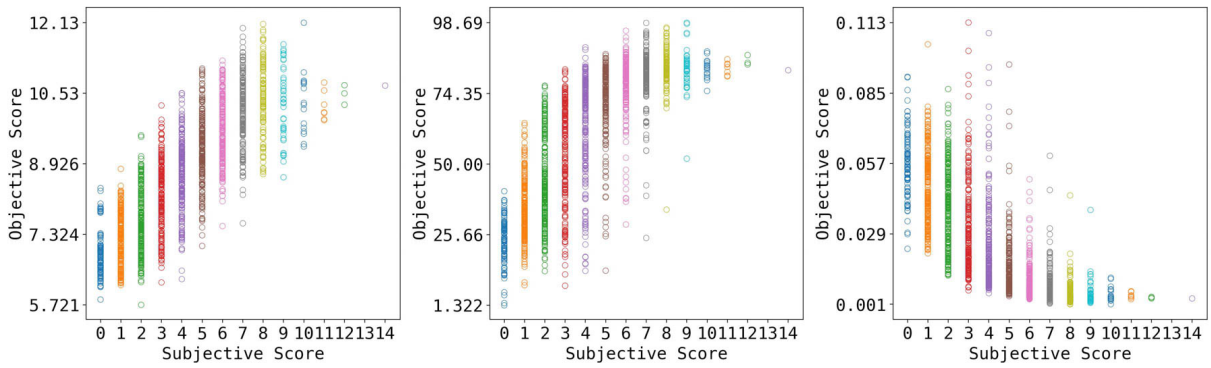


(h) ResNet-50 [64]

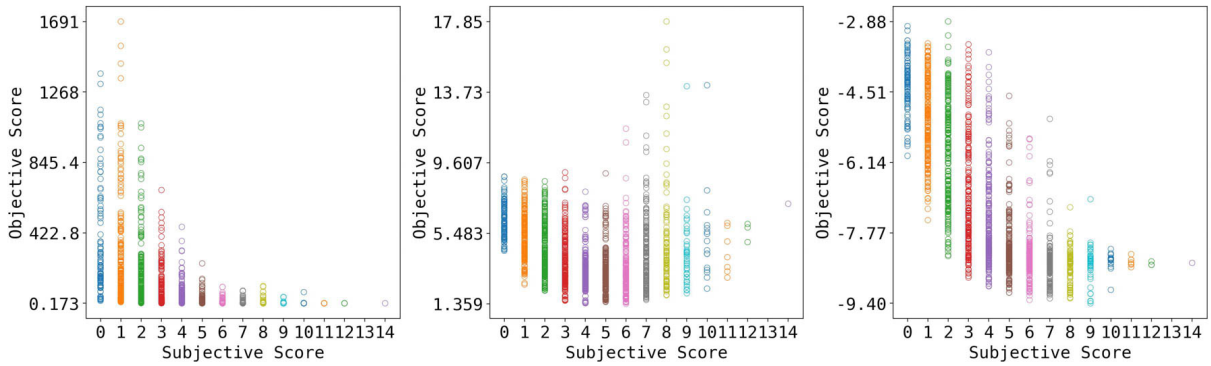


(i) ResNet-101 [64]

Figure 3.19: Scatter plots of absolute z-level versus predicted scores on the FocusPath dataset.

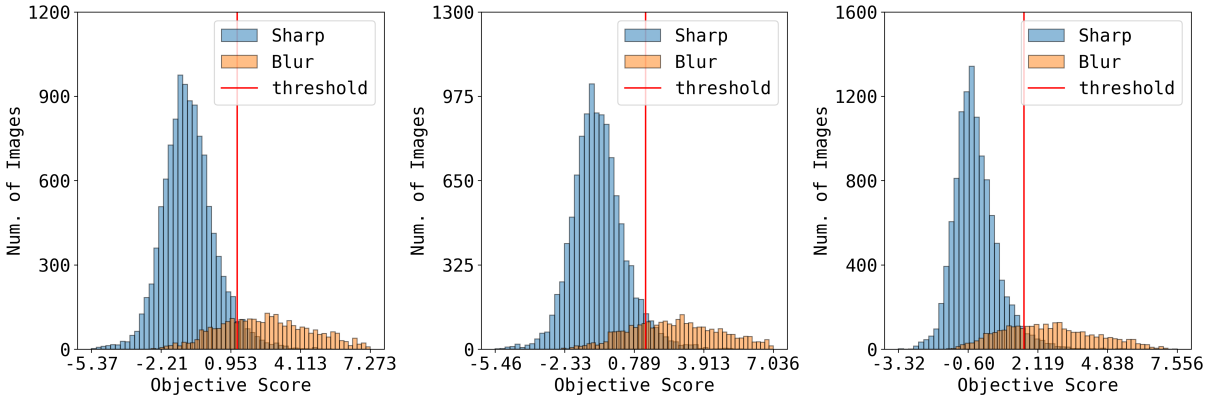


(j) FQPath [57] (k) Synthetic MaxPol [58] (l) LPC [107]

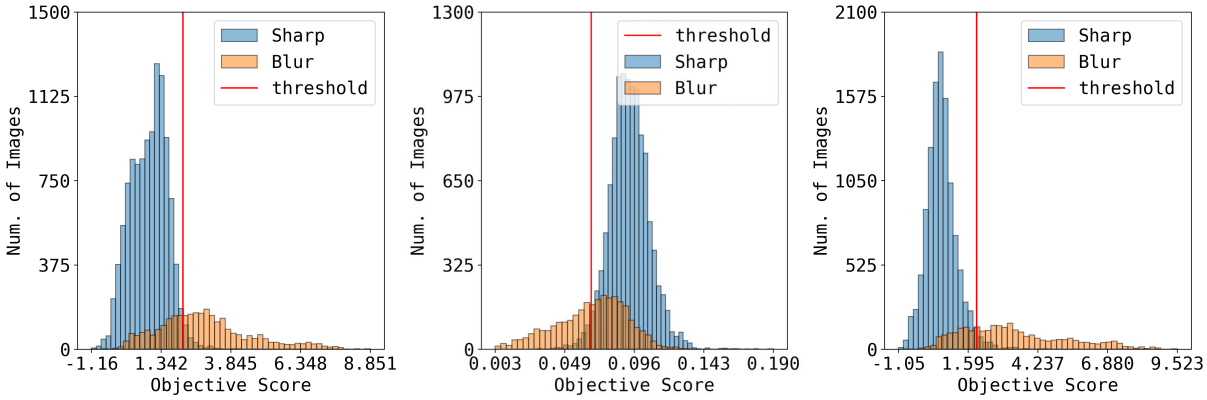


(m) GPC [108] (n) SPARISH [116] (o) HVS-MaxPol-1 [2]

Figure 3.19: Scatter plots of absolute z-level versus predicted scores on the FocusPath dataset (continued).

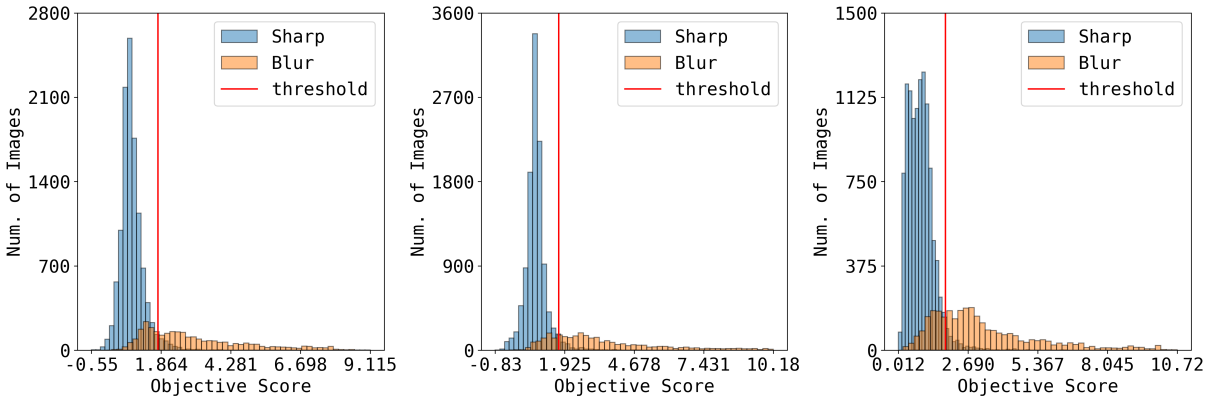


(a) FocusLiteNN (1-kernel) (b) FocusLiteNN (2-kernel) (c) FocusLiteNN (10-kernel)

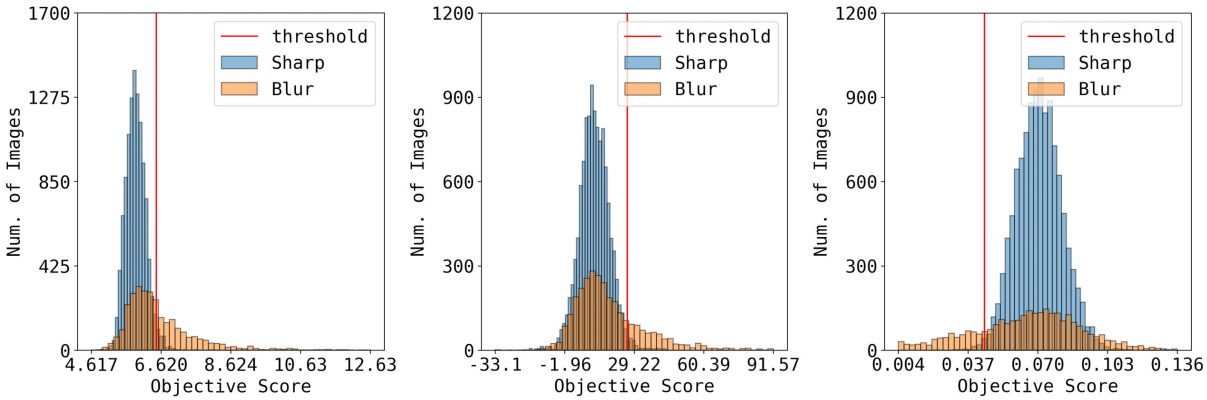


(d) EONSS [174] (e) MLV [93] (f) DenseNet-13

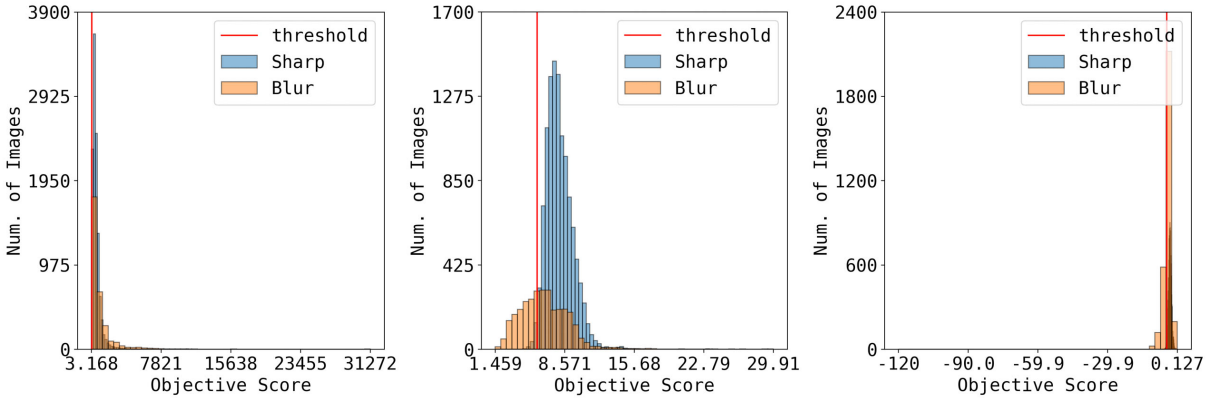
Figure 3.20: Histogram of objective scores on the TCGA@Focus dataset.



(g) ResNet-10 (h) ResNet-50 [64] (i) ResNet-101 [64]



(j) FQPath [57] (k) Synthetic MaxPol [58] (l) LPC [107]



(m) GPC [108] (n) SPARISH [116] (o) HVS-MaxPol-1 [2]

Figure 3.20: Histogram of objective scores on the TCGA@Focus dataset (continued).

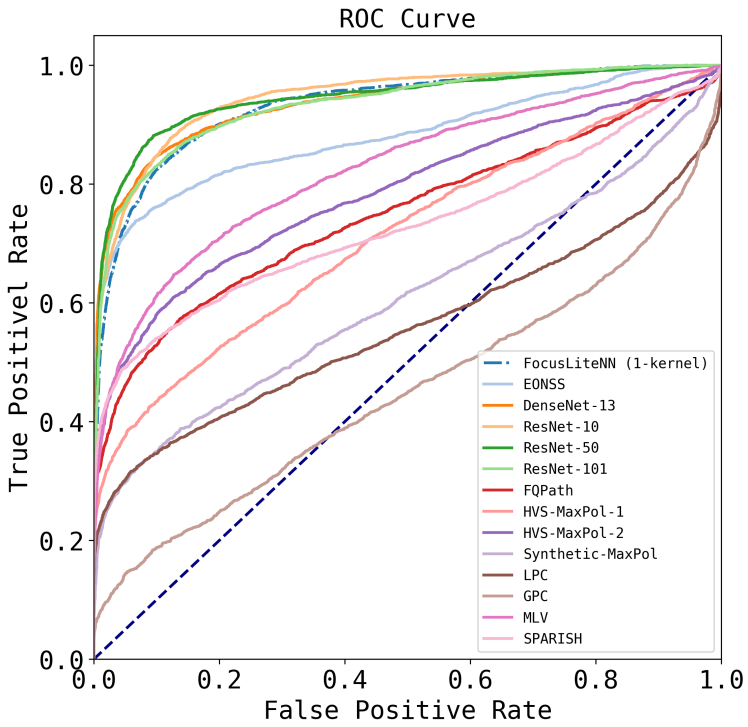
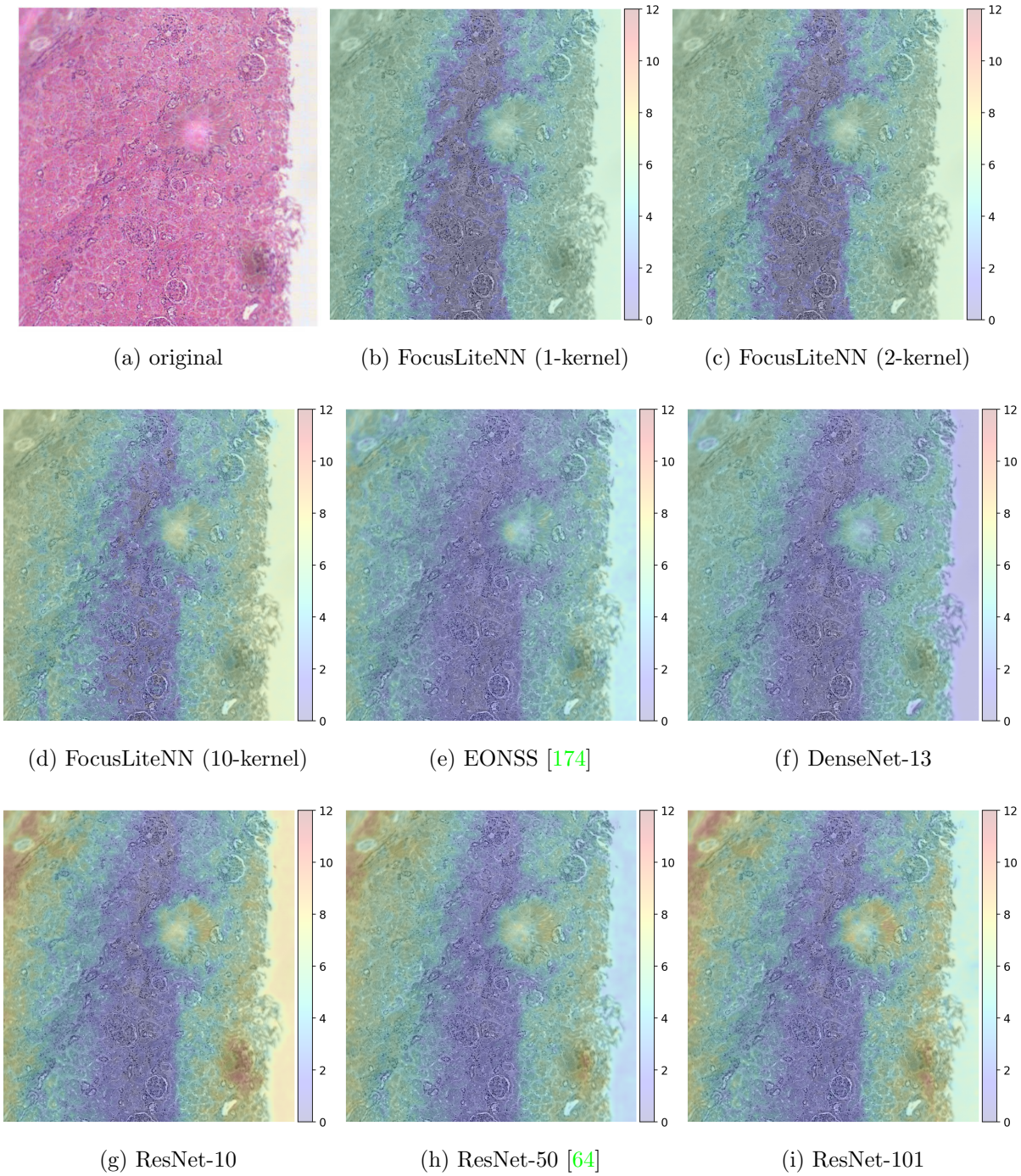


Figure 3.21: ROC curves of the testing models evaluated on the TCGA@Focus dataset.

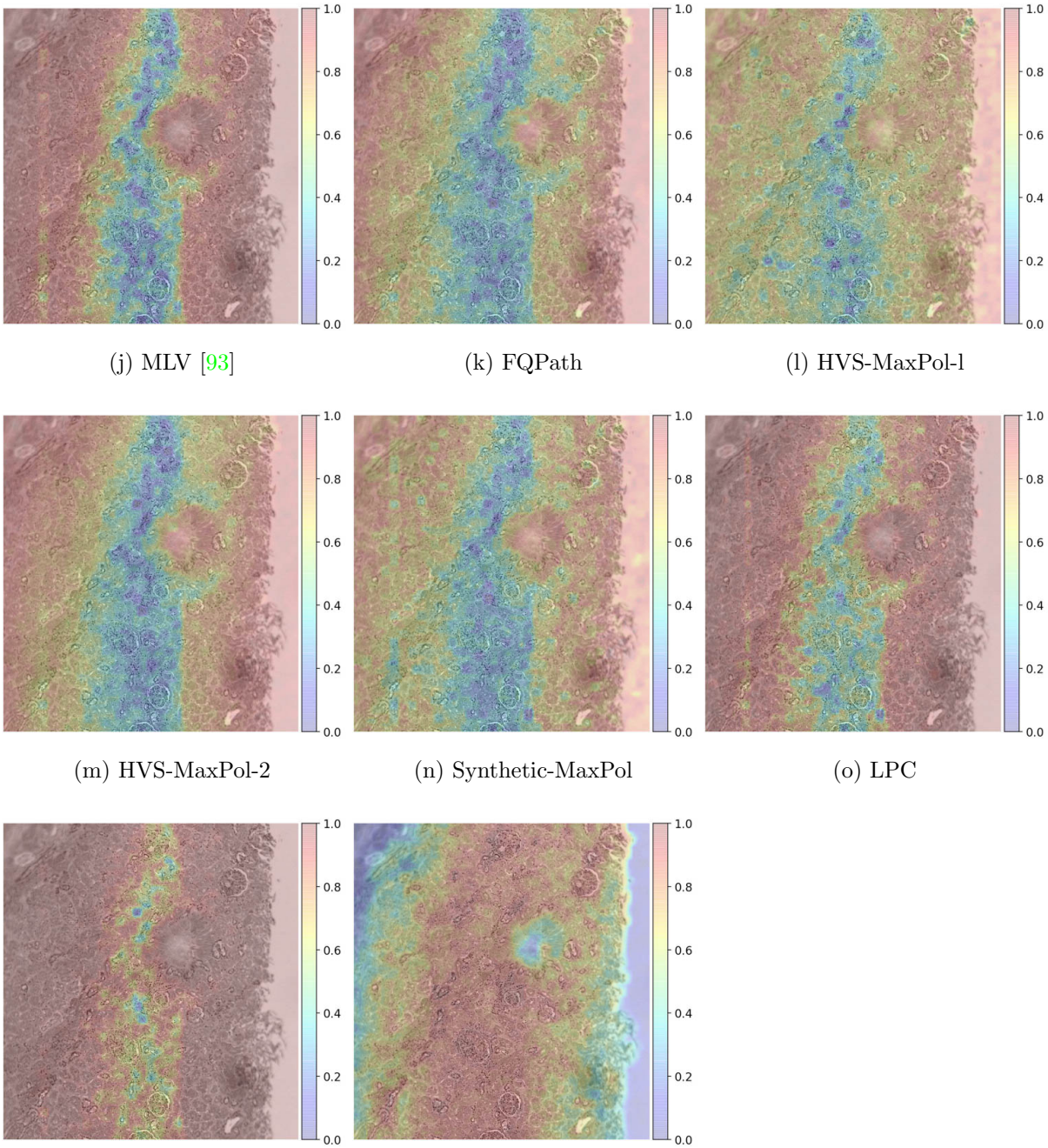


(a) original (b) FocusLiteNN (1-kernel) (c) FocusLiteNN (2-kernel)

(d) FocusLiteNN (10-kernel) (e) EONSS [174] (f) DenseNet-13

(g) ResNet-10 (h) ResNet-50 [64] (i) ResNet-101

Figure 3.22: Absolute heatmaps. A higher score indicates more blurriness. The predicted scores represent the z-levels in the FocusPath dataset.

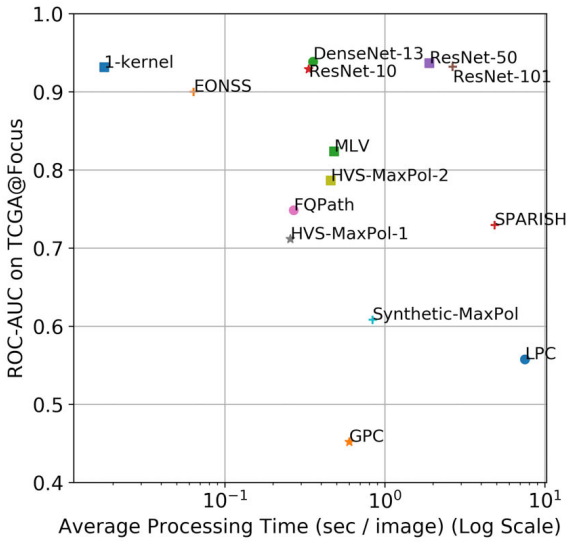


(j) MLV [93] (k) FQPath (l) HVS-MaxPol-l

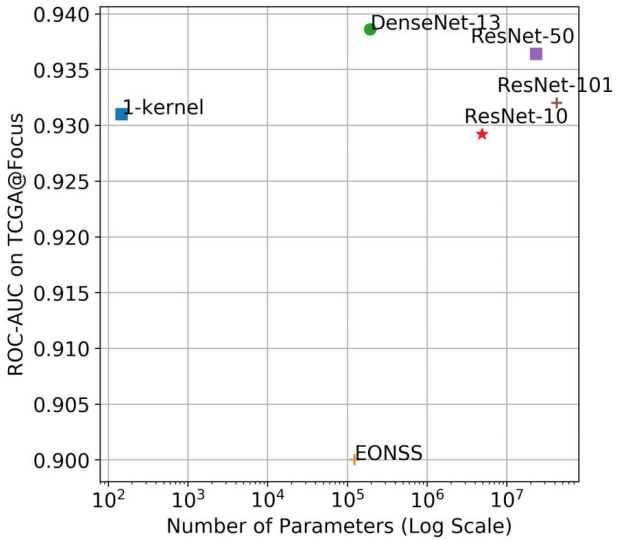
(m) HVS-MaxPol-2 (n) Synthetic-MaxPol (o) LPC

(p) GPC (q) SPRISH

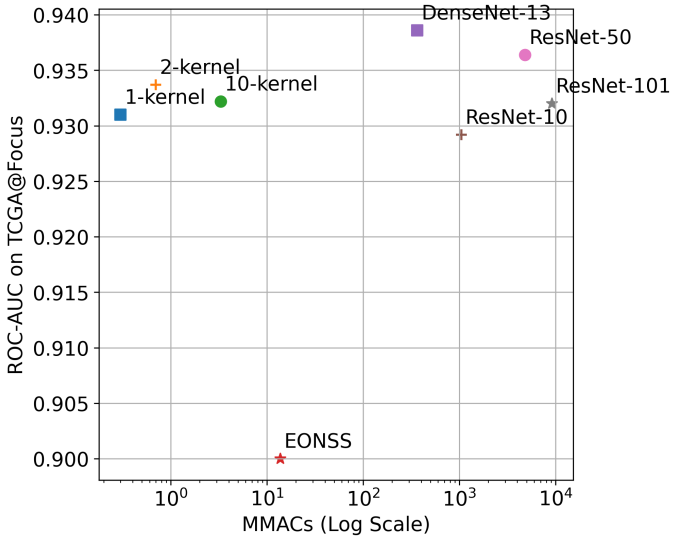
Figure 3.22: Normalized heatmaps. A higher score indicates more blurriness. The predicted scores of each model are independently linearly normalized to the range 0 to 1. (continued).



(a) ROC-AUC v.s. CPU Time



(b) ROC-AUC v.s. # Model Params



(c) ROC-AUC v.s. # MMACs

Figure 3.23: Average processing time versus ROC-AUC, model size versus ROC-AUC and MMACs versus ROC-AUC on the TCGA@Focus Dataset. The x-axis of each figure is on a log scale. All models are running on an Intel i9-7920X @ 2.90GHz with 32 GB memory. ROC-AUC: Area under the receiver operating characteristic curve. MMACs: Million multiply-accumulate operations.

Chapter 4

Unsupervised IQA Score Fusion by Deep Maximum a Posteriori Estimation

4.1 Introduction

IQA models are designed to predict the perceptual quality of images. Over the past decades, numerous IQA models have been introduced. Their correlation with human evaluations, typically represented by the MOS, has progressively increased. Many IQA models demonstrate superior performance on average when being evaluated on specific datasets. However, due to their distinct design philosophies and implementation details, they often capture some particular types of distortions or handle some specific image contents better than others. Consequently, individual IQA models often fail to address all types of images and distortions encountered in real-world scenarios. An intuitive idea to quickly boost IQA performance without developing a new one is to leverage existing IQA models by fusing their scores to attain more reliable predictions, so as to harness the strengths and mitigate the weaknesses in each model. Using score fusion techniques, we can synthesize MOS for large-scale datasets where conducting subjective testing is challenging [178].

Existing fusion approaches can be categorized non-mutually exclusively into empirical, rank fusion methods, and supervised learning-based [125]. Empirical models [126, 127, 128, 129, 130, 131, 132, 133] fuse a predetermined set of IQA models using a handcrafted formula. This approach significantly constrains its adaptability when introduced with new IQA models. Rank fusion methods operate in the discrete rank domain, where the range of all IQA models is mapped to the same uniform distribution. However, these methods are tied closely to the diversity of the ranking dataset, which can impede generalizability. Supervised learning-based methods [134, 135] are trained under the guidance of the MOS of a single subjective rated dataset. Such fusion methods are essentially refined versions of supervised learning-based IQA models since they share the same ground truth, i.e., MOS of a specific dataset, as the base IQA models. These fusion methods are more closely related to supervised learning-based IQA models since they share the same ground truth, MOS. Nevertheless, these black-box models often suffer from limited generalizability and lack of explainability.

We argue that supervised learning-based fusion approaches, i.e., when MOS is used as the guidance, in essence, counter the fundamental reasoning behind the score fusion idea in its attempts to improve generalizability, because any MOS is associated with some specific image content and specific distortion types and levels, and thus inevitably leads to biases that we try to avoid. With the guidance of MOS, one can easily evaluate the performance of each model in the fusing list. This allows one to choose the most effective models to fuse. Without MOS as a reference, it becomes difficult to justify the effectiveness of each model, making IQA score fusion inherently a different and more challenging task. In this work, we focus on unsupervised score fusion *without* involving MOS. We believe that the key to the problem is to estimate the uncertainty of each IQA score without any ground truth. Existing fusion methods either lack uncertainty estimation or only remain coarse-grained at the model level. To mitigate this problem, we proposed a general framework for unsupervised IQA score fusion using MAP estimation. Our framework consists of an encoder, a set of decoders and a set of uncertainty estimation modules. The encoder fuses individual IQA models, which not only expedites inference but also enhances the framework's explainability. The decoders model the relationship between MOS and each IQA model. Uncertainty estimation modules are responsible for fine-grained, score-level

uncertainty estimation. Given its unsupervised nature, the proposed model demonstrates superior generalizability to unseen data. The fine-grained uncertainty estimation module predicts the uncertainty of each score generated by individual IQA models. By identifying and fusing the less uncertain portions of each model, we achieve a more accurate and reliable prediction. The proposed method is trained end-to-end to ensure all modules collaborate seamlessly. An overview of the framework is shown in Fig. 4.1. Comprehensive experiments on ten testing sets demonstrate the superiority of the proposed model over other ones.

The main novelties of our work include:

1. To the best of our knowledge, we propose the first unsupervised learning-based score fusion approach for IQA.
2. We formalize the first observation model of IQA fusion and address the task using MAP estimation.
3. By building a powerful fine-grained uncertainty estimation module, the proposed model increases accuracy and reduces uncertainty in its prediction by harnessing the strengths and mitigating the weaknesses of each model.
4. We show that rank fusion can be easily integrated into our general framework.
5. The proposed model exhibits the capability of rejecting "bad" models in the fusion process.

4.2 Proposed Framework

4.2.1 Observation Model

Given N distorted images $\{I_i^d | i = 1, 2, \ldots, N\}$ and their corresponding pristine ones $\{I_i^r | i = 1, 2, \ldots, N\}$, we can evaluate the quality of the distorted images using a Full Reference (FR) IQA metric: $x_i^j = \textit{FR-IQA}_j(I_i^d, I_i^r)$ or a No Reference (NR) one: $x_i^j = \textit{NR-IQA}_j(I_i^d)$. The

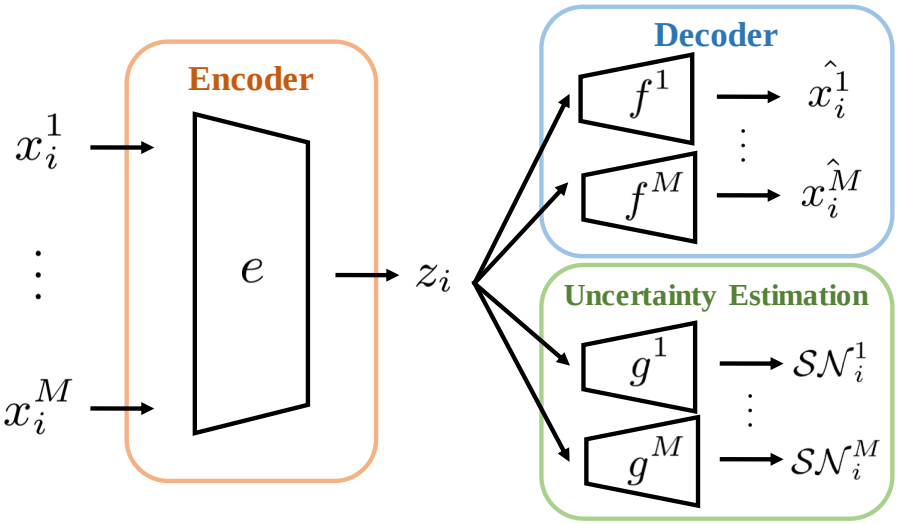


Figure 4.1: The diagram of the proposed framework fusing IQA scores $\mathbf{x_i} = \{x_i^j | j = 1, 2, ..., M\}$ of image I_i. $\hat{x_i^j} = f^j(z_i)$ represents the reconstructed score. $\mathcal{SN}_i^j$ is the abbrev. of $\mathcal{SN}(\xi_i^j, \omega_i^j, \alpha_i^j)$, which is the predicted conditional Skew Normal distribution $p(x_i^j | z_i)$.

proposed framework is capable of fusing IQA metrics of all kinds, either FR-IQA or NR-IQA. Suppose that we have scores $\{x_i^j \in \mathbb{R} | i = 1, 2, \ldots, N; j = 1, 2, \ldots, M\}$ generated using M IQA metrics. x_i^j is the score of the i^{th} image evaluated by the j^{th} metric. We adopt the notation where subscripts represent images and superscripts denote models. Our goal is to fuse these scores into final predictions $\{z_i \in \mathbb{R} | i = 1, 2, \ldots, N\}$ that align with MOS better than any individual IQA metric. We assume that the score x_i^j is generated by a function of z_i and two independent noise: a score-dependent noise n_i^j and a model-dependent noise $\hat{n^j}$:

$$x_i^j = f^j(z_i) + n_i^j + \hat{n^j} \tag{4.1}$$

where $f^j : \mathbb{R} \to \mathbb{R}$ is a deterministic function that simulates the mapping from MOS to score x_i^j. n_i^j is a score-dependent noise that follows a skew normal distribution: $n_i^j \sim \mathcal{SN}(\xi_i^j, \omega_i^j, \alpha_i^j)$ where ξ_i^j, ω_i^j, α_i^j are the location, scale and shape parameters, respectively. ω_i^j is determined through a score-level uncertainty estimation function $g^j : \mathbb{R} \to \mathbb{R}$: $\omega_i^j =$

$g^j(z_i)$. For simplicity, we assume $\xi_i^j = 0$ and α_i^j is only model-dependent, therefore we denote it as α^j throughout the paper. $\hat{n^j}$ is a model-dependent noise that follows a zero mean normal distribution: $\hat{n^j} \sim \mathcal{N}(0, \sigma^{j^2})$. Since n_i^j and $\hat{n^j}$ are independent, we will show later that $n_i^j + \hat{n^j}$ also follows a skew normal distribution. We assume f^j and g^j follow some parametric forms with parameters θ_{f^j} and θ_{g^j}, respectively.

The motivation behind modeling uncertainty as score-dependent arises from the human visual system's increased uncertainty when evaluating lower-quality images. Similar behavior is also found in many IQA models, showing greater variance in regions of lower or middle MOS ranges. As a result, it is more accurate to model the uncertainty as score-dependent. This is demonstrated in the variance of the fitted conditional density curves shown in Fig. 4.2. The reason for modeling the conditional distribution $p(x_i^j|z_i)$ as asymmetric is also based on the observation across various IQA models: $p(x_i^j|z_i)$ is usually skewed towards the lower score side. We choose skew normal due to computational feasibility. This characteristic is also illustrated in the fitted conditional density curves in Fig. 4.2, where we plot the empirical distributions of IWSSIM [4], FSIM [5], VSI [6] evaluated on the KADID-10K [7] and VCLFER [8] dataset.

4.2.2 MAP Formulation and Optimization

Our method is based on the Maximum a Posteriori (MAP) estimation framework. It's easy to show that the likelihood $p(x_i^j|z_i) = p(n_i^j + \hat{n^j}|z_i)$ is also a skew normal distribution $\mathcal{SN}(0, \tilde{\omega_i^j}, \tilde{\alpha^j})$. The likelihood function can be written as

$$
\begin{aligned}
& p(x_i^j|z_i; \tilde{\alpha^j}, \tilde{\omega_i^j}) \\
=& p(x_i^j - z_i|z_i; \alpha^j, \omega_i^j, \sigma^j) \\
=& p(x_i^j - f^j(z_i)|z_i; \theta_{f^j}, \alpha^j, \theta_{g^j}, \sigma^j) \\
=& \frac{\sqrt{2}}{\sqrt{\pi}\tilde{\omega_i^j}} \left(\frac{1}{2} + \frac{\operatorname{erf}\left(\frac{\tilde{\alpha^j}\left(x_i^j - f^j(z_i)\right)}{\sqrt{2}\tilde{\omega_i^j}} \right)}{2} \right) e^{-\frac{(x_i^j - f^j(z_i))^2}{2\tilde{\omega_i^j}}}
\end{aligned}
\tag{4.2}
$$

where $\tilde{\omega_i^j} = \sqrt{{\omega_i^j}^2 + {\sigma^j}^2} = \sqrt{g^j(z_i)^2 + {\sigma^j}^2}$ and $\tilde{\alpha^j} = \frac{\alpha_i^j \omega_i^j}{\sqrt{{\omega_i^j}^2 + {\sigma^j}^2 + {\alpha^j}^2 {\sigma^j}^2}} = \frac{\alpha_i^j g^j(z_i)}{\sqrt{g^j(z_i)^2 + {\sigma^j}^2 + {\alpha^j}^2 {\sigma^j}^2}}$ are the new scale and shape parameters, respectively. erf is the error function defined as $\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} \, \mathrm{d}t$. The MAP objective can be written as

$$
\begin{aligned}
& \underset{\mathbf{z}, \boldsymbol{\alpha}, \boldsymbol{\sigma}, \boldsymbol{\theta_f}, \boldsymbol{\theta_g}}{\arg\max} \; p(\mathbf{z}|\mathbf{X}; \boldsymbol{\alpha}, \boldsymbol{\sigma}, \boldsymbol{\theta_f}, \boldsymbol{\theta_g}) \\
&= \underset{\mathbf{z}, \boldsymbol{\alpha}, \boldsymbol{\sigma}, \boldsymbol{\theta_f}, \boldsymbol{\theta_g}}{\arg\max} \; \log p(\mathbf{X}|\mathbf{z}; \boldsymbol{\alpha}, \boldsymbol{\sigma}, \boldsymbol{\theta_f}, \boldsymbol{\theta_g}) + \log p(\mathbf{z}) \qquad (4.3) \\
&= \underset{\substack{z_i, \alpha^j, \sigma^j \\ \theta_{f^j}, \theta_{g^j}}}{\arg\max} \; \sum_j^M \sum_i^N \left(\log p(x_i^j | z_i; \alpha^j, \sigma^j, \theta_{f^j}, \theta_{g^j}) + \log p(z_i) \right) \qquad (4.4)
\end{aligned}
$$

where $\mathbf{X} \in \mathbb{R}^{N \times M}$ represents the IQA scores of the N images calculated by M IQA models. $\mathbf{z} \in \mathbb{R}^N$ is the MOS of the corresponding N images. $\boldsymbol{\alpha} \in R^M$, $\boldsymbol{\sigma} \in R^M$ are the collections of distributional parameters α^j, σ^j, respectively. $\boldsymbol{\theta_f}, \boldsymbol{\theta_g}$ are the collections of model parameters of f^j, g^j, respectively. $p(\mathbf{z})$ is the prior distribution over the predicted MOS $\mathbf{z}$. To make Eq. 4.3 tractable, we introduce two commonly made assumptions. First, IQA models are conditionally independent, i.e., $p(\mathbf{x_i}|z_i) = \prod_{j=1}^M p(x_i^j|z_i)$ where $\mathbf{x_i} \in \mathbb{R}^M$ are the scores of the same image evaluated by M IQA metrics. Second, the images are independent, that is, $p(\mathbf{X}) = \prod_{i=1}^N p(\mathbf{x}_i)$. Now we can factorize Eq. 4.3 into a summation of individual loss functions as shown in Eq. 4.4. By substituting Eq. 4.2 into Eq. 4.4, we get the final objective. The prior distribution $p(\mathbf{z})$ describes our understanding of the dataset. However, given the variability of MOS distributions across datasets, the use of informative priors might compromise generalizability. Without loss of generality, we define $p(\mathbf{z}) = \mathcal{U}(0, 1)$ since most subjective rated datasets advertise themselves as diverse where the MOS is close to uniform in practice [125]. This prior is used to regularize the range of the predicted $\mathbf{z}$ without compromising generalizability. Finally, we jointly optimize $\mathbf{z}, \boldsymbol{\alpha}, \boldsymbol{\sigma}, \boldsymbol{\theta_f}, \boldsymbol{\theta_g}$ end-to-end.

4.2.3 Amortized Inference

In the original MAP formulation in Eq. 4.3, the estimation of $\mathbf{z}$ is carried out iteratively on a dataset. This poses two challenges: it may not yield an accurate estimation when the scores of a singular image are available, and it tends to be time-consuming. To address these limitations, we introduce amortized inference, which simplifies the inference procedure without stressing the training. Instead of optimizing $\mathbf{z}$ directly, an encoder $e : \mathbb{R}^M \rightarrow \mathbb{R}$ is introduced to fuse the scores of a given image. The encoder can be parameterized as a Fully Connected Network (FCN) with weights θ_e. An overview of the framework is shown in Fig. 4.1. During inference, the framework no longer needs a large dataset to do optimization. Instead, it requires only the M scores of the testing image: $z_i = e(\mathbf{x_i}; \theta_e)$. By substituting z_i with $e(\mathbf{x_i}; \theta_e)$ in Eq. 4.4, the encoder is jointly optimized with other parameters. The amortized objective now becomes

$$\underset{\boldsymbol{\alpha}, \hat{\boldsymbol{\sigma}}, \boldsymbol{\theta_f}, \boldsymbol{\theta_g}, \theta_e}{\arg\max} \log p(\mathbf{X} | e(\mathbf{X}; \theta_e); \boldsymbol{\alpha}, \hat{\boldsymbol{\sigma}}, \boldsymbol{\theta_f}, \boldsymbol{\theta_g}) + \log p(e(\mathbf{X}; \theta_e)) \tag{4.5}$$

4.2.4 Rank Fusion

Rank fusion offers a non-linear transformation of diverse score distributions to a consistent uniform distribution. Such a transformation helps to stabilize the fusion process. In this subsection, we show that rank fusion can be seamlessly integrated into our general framework. Rather than optimizing in the discrete space which is computationally intensive, we opt for a more effective solution by normalizing the discrete ranks and conducting our optimization in the continuous space.

We compute the normalized rank $r_i^j \in \mathbb{R}$ corresponding to the original score x_i^j by finding the index $R_i^j \in \mathbb{N}$ of x_i^j in the ascendingly ranked $\mathbf{x^j} = \{x_i^j | i = 1, 2, ..., N\}$. Subsequently, r_i^j is derived by $r_i^j = R_i^j / N$. Similar to Fig 4.2, we show in Fig 4.3 the empirical distribution of the normalized rankings of IWSSIM [4] and FSIMc [5] evaluated on the KADID-10K [7] dataset. It is easy to find the same observation model that can be applied to the rank case without modification.

By replacing x_i^j with r_i^j in Eq. 4.4, we formulate rank fusion as a special instance of the general framework. The objective of rank fusion can be written as

$$\underset{\substack{z_i,\,\alpha^j,\,\sigma^j\\ \theta_{f^j},\,\theta_{g^j}}}{\arg\max} \sum_{j}^{M}\sum_{i}^{N}\left(\log p(r_i^j|z_i;\alpha^j,\sigma^j,\theta_{f^j},\theta_{g^j})+\log p(z_i)\right) \tag{4.6}$$

The benefit of rank fusion is that the mappings $\mathbf{f}=\{f^j;j=1,2,...M\}$, which represent the characteristics of various IQA metrics, are being transformed into a similar form with the same range. This simplifies the estimation of $\mathbf{f}$. Such a transform potentially stabilizes the optimization process, especially when integrating extreme IQA models with atypical $\mathbf{f}$ mappings.

4.3 Experiments

4.3.1 Implementation and Experimental Details

To rigorously assess the fusion performance of the proposed method, we evaluate it on ten diverse subjectively rated IQA datasets. Six of them, LIVE R2 [190], TID2013 [191], CSIQ [192], VCL@FER [8], CIDIQ50 and CIDIQ100 [193], feature single distortion. Four others, MDID [194], MDID2013 [195], LIVE MD [196] and MDIVL [197] comprise multiply distorted images. To demonstrate the generalizability of the proposed method, we include 12 FR-IQA and 4 NR-IQA methods of diverse design philosophies [198] and varying correlation w.r.t. MOS (see Table 4.1). The chosen metrics include both traditional methods and deep learning-based methods. The metrics are VSI [6], FSIMc [179], IWSSIM [4], DSS [181], MCSD [183], CID MS [184], GMSD [185], FSIM [5], SFF [180], QASD [182], VIF [47], VIF DWT [186], HOSA [187], NIQE [50], MEON [188], and QAC [189]. The goal of this paper is to develop an IQA score fusion model, rather than a new IQA model. So we do not include individual IQA metrics outside of the fusion list in comparison. Nonetheless, our framework remains extendable, allowing for seamless integration of state-of-the-art metrics to enhance fusion results. Except for empirical fusion methods that use a predetermined set of IQA metrics, we use the same set of 16 metrics for all other fusion methods in all of our experiments. RRF [144, 143] and RRFW [145, 146] are the only two unsupervised

learning-based IQA fusion models in the literature. We refer to RRF as "unsupervised learning-based," where the training set is used to calculate the ranked index of the testing image. For a fair comparison, unsupervised training-based methods are retrained on the large KADID10K [7] dataset and tested on the ten testing sets mentioned above. Supervised learning-based fusion methods are not included in the comparison since the main focus of the paper is unsupervised fusion. To evaluate the performance, Spearman's Rank Correlation Coefficient (SRCC) and Pearson Linear Correlation Coefficient (PLCC) over the ten testing sets are provided in Table 4.1.

Our framework is versatile, allowing for diverse implementations of the encoder e, decoder f and uncertainty estimation module g. The chosen implementations are used to demonstrate the general framework and are not necessarily the optimal configurations. For simplicity, we implement the encoder e using a six-layer Fully Connected Network (FCN). The number of input and output channels of each layer is set to the number of models to be fused. LeakyReLU is added to each layer except for the last one. Finally, for each testing image, the FCN is used to predict a set of weights adaptive to the content to combine the scores linearly. This makes the encoder explainable since we can inspect the predicted weights. Due to its lightweight nature, the inference speed is generally very fast. We have tried using the same Cascade Neural Network in CNNM [142] as the encoder, which also gives similar results. Since the goal of most IQA metrics is to approximate human perception of quality which is measured by MOS or DMOS (Difference of MOS), it's natural to assume f^j is generally monotonic w.r.t. the predicted MOS z_i. We implement f^j as an exponential function in the form: $f^j(z_i^j) = -e^{a^j(z_i^j - b^j)} + c^j$. We have also tried using a Cascade Neural Network or a MLP, which results in similar performance. The uncertainty estimation function g^j is chosen to be quadratic in the form: $g^j(z_i^j) = a^j {z_i^j}^2 + b^j z_i^j + c^j$. We train the parameters in f and g with other modules end-to-end using the Adam optimizer with a learning rate of 0.002. We stop the training when the loss reaches a plateau.

4.3.2 Evaluation Results

We introduce three variants of the proposed model: SF-ms, RF-ms, and SF-m. "SF" stands for Score Fusion while "RF" denotes Rank Fusion. The suffixes further detail

the uncertainty estimation levels: "-m" is Model-level while "-s" stands for Score-level. "-ms" means both levels of estimations are included. Table 4.1 and Table 4.2 provide a comprehensive evaluation of the proposed model, seven other fusion methods, and 16 individual IQA models used in the fusion across ten diverse testing datasets. Here we also include two supervised learning-based fusion methods: MMF [134] and CNNM [142]. The proposed models with fine-grain uncertainty estimation even outperform other fusion methods, as well as individual IQA metrics, in terms of average SRCC and PLCC. The supervised learning-based method CNNM performs slightly inferior than SF-ms and RF-ms, we hypothesize that this is due to the domain shift between the training and testing datasets.

A statistical significance test is also provided in Table 4.3. Each entry in Table 4.3 is composed of ten symbols, with each representing the result of the statistical significance testing between two models on one IQA dataset. The order of the database is the same as in Table 4.1. 1 means that the method is statistically better than the method in the column on that particular dataset with 95% confidence, 0 means that it is statistically worse, and - means that it is statistically indistinguishable. Due to the page limit, we only include one individual IQA model, IWSSIM, which has the highest average SRCC and PLCC as a reference point. It can be shown that all versions of the proposed model are significantly better than RRF, RRFW and MMF on all testing datasets. The proposed model also outperforms IWSSIM on the majority of the testing datasets. CNNM performs on par with SF-ms and RF-ms, which makes sense since CNNM is a supervised learning-based method while the proposed methods are unsupervised.

Owing to the fine-grained score-level uncertainty estimation, SF-ms and RF-ms rank among the top three in five and six out of ten individual testing sets, respectively. To further demonstrate this, an ablation study is conducted where the score-level uncertainty estimation is removed, resulting in the SF-m model. As anticipated, SF-m is inferior to SF-ms, although it surpasses other fusion methods on average, thanks to the accurate estimation of f and model-level uncertainty. The predicted overall scale parameter $\tilde{\omega^j} \in \mathbb{R}^M$ of each IQA model is also negatively correlated to the SRCC and PLCC of individual models. Since the proposed framework can easily be extended to rank fusion, we include RF-ms in comparison. This also provides a fair testing environment for other fusion-based

methods, such as RRF and RRFW, since the scores are converted to ranks in the same way. As shown in the table, RF-ms outperforms other rank fusion methods, demonstrating the superiority of the proposed framework.

To further demonstrate the importance of both model-level and score-level uncertainty estimation, we carried out an additional experiment where two "bad-performing" IQA metrics are included in the fusing list. Without any prior knowledge of these metrics, many unsupervised learning-based methods may suffer since distinguishing "bad" metrics becomes challenging without MOS or precise uncertainty estimation. We implement the two "bad" metrics using random number generators producing uniformly sampled numbers from 0 to 1, which is the most common range for FR-IQA models. The results are shown in Table 4.4. The performances of RRF and RRFW decline significantly in this case, while the performance of the proposed method remains competitive. A closer examination of the uncertainty estimation module reveals that the estimated overall scale parameters $\tilde{\omega^j}$ of the two "bad" metrics are significantly higher than others. As a result, the encoder e assigns very little weight to these two metrics.

One drawback of supervised learning-based methods is that they are sensitive to the "quality" of the training data, i.e., the accuracy of the ground truth MOS. We demonstrate in Table 4.5 that when the ground truth of the training data is contaminated by noise, the performance of supervised learning-based methods will drop significantly. The training data and testing data are kept the same across all models and also the same as in Table 4.1, expects that the MOS of the training data is altered. We synthesize the contaminated training data by adding zero-mean Gaussian noise to the MOS, with a standard deviation of $0.3 \cdot \text{std}$, where std is the standard deviation of the MOS. It can be shown that the performance of both MMF and CNNM drops significantly compared to the results in Table 4.1. However, due to the unsupervised nature, the proposed model will not be affected.

Similarly, when the number of samples is very limited in the training dataset, the supervised learning-based methods will overfit and result in poor performance on the testing dataset. For rank fusion methods, the rankings will be less reliable due to the limited and less diverse samples. Consequently, their performance will also decline. To demonstrate this, we form a new training dataset by randomly sampling 100 images and their corresponding MOS from the KADID10k dataset. We retrain all models on this subset and

test them on the same testing as in Table 4.1. According to the results in Table 4.6, the performance of RRF, RRFW, MMF and CNNM all drop significantly compared to the one in Table 4.1.

FQA Score Fusion

Beyond general IQA score fusion, our proposed framework is versatile and can be applied to a wide range of quality assessment tasks. For instance, WSI FQA, which evaluates the out-of-focus level in microscopic WSIs, is a specific case of IQA. We conducted an experiment fusing seven diverse FQA models: FQPath [57], HVS-MaxPol-1 [2], HVS-MaxPol-2 [2], Synthetic-MaxPol [58], LPC [107], GPC [108], and SPARISH [116]. We compared the performance of both rank fusion methods (RRF [144] and RRFW [145]) and supervised fusion methods (MMF [134] and CNNM [142]). Due to FocusPath [2] being the only publicly available dataset that contains continuous labels and large patches, we divided the dataset into 50% training and 50% testing sets. All trainable fusion models were trained from scratch using the training set. The evaluation results on the testing set are summarized in Table 4.7. Unlike general IQA score fusion, the supervised method CNNM [142] stands out among the other fusion methods. This is attributed to the fact that training and testing are performed on the same dataset, eliminating any domain gap. Consequently, it is expected that supervised fusion methods would achieve better performance. However, since the domain gap is more significant in the IQA fusion case, the performance of the supervised methods will deteriorate. Despite this, our proposed framework still outperforms all other fusion methods, including individual FQA models.

We conducted a comparison of the inference speed of various fusion methods, with results presented in Table 4.8. This analysis focuses solely on the time required for the fusion process, excluding the time spent calculating the IQA scores. Empirical fusion methods were excluded from this comparison due to their dependence on predefined sets of IQA models. During inference, rank fusion methods require ranking scores across the entire dataset, making their speed sensitive to dataset size. RRFW [145] requires iterative optimization during inference, which further decreases the speed. In this experiment, the dataset comprises 10,000 images, each evaluated by 16 IQA models. MMF [134] employs

an SVR encoder, which generally exhibits slower inference speeds compared to neural network-based encoders such as the cascade neural network utilized in CNNM [142]. For our proposed framework, we tested three encoders: the default MLP-based encoder, a cascade neural network encoder identical to the one in CNNM [142], and a linear model. Our results indicate that the proposed framework, equipped with the default MLP encoder, outperforms both supervised and rank fusion methods in terms of inference speed. Given the flexibility of our fusion framework to accommodate various encoders, further speed improvements are possible by using a linear encoder.

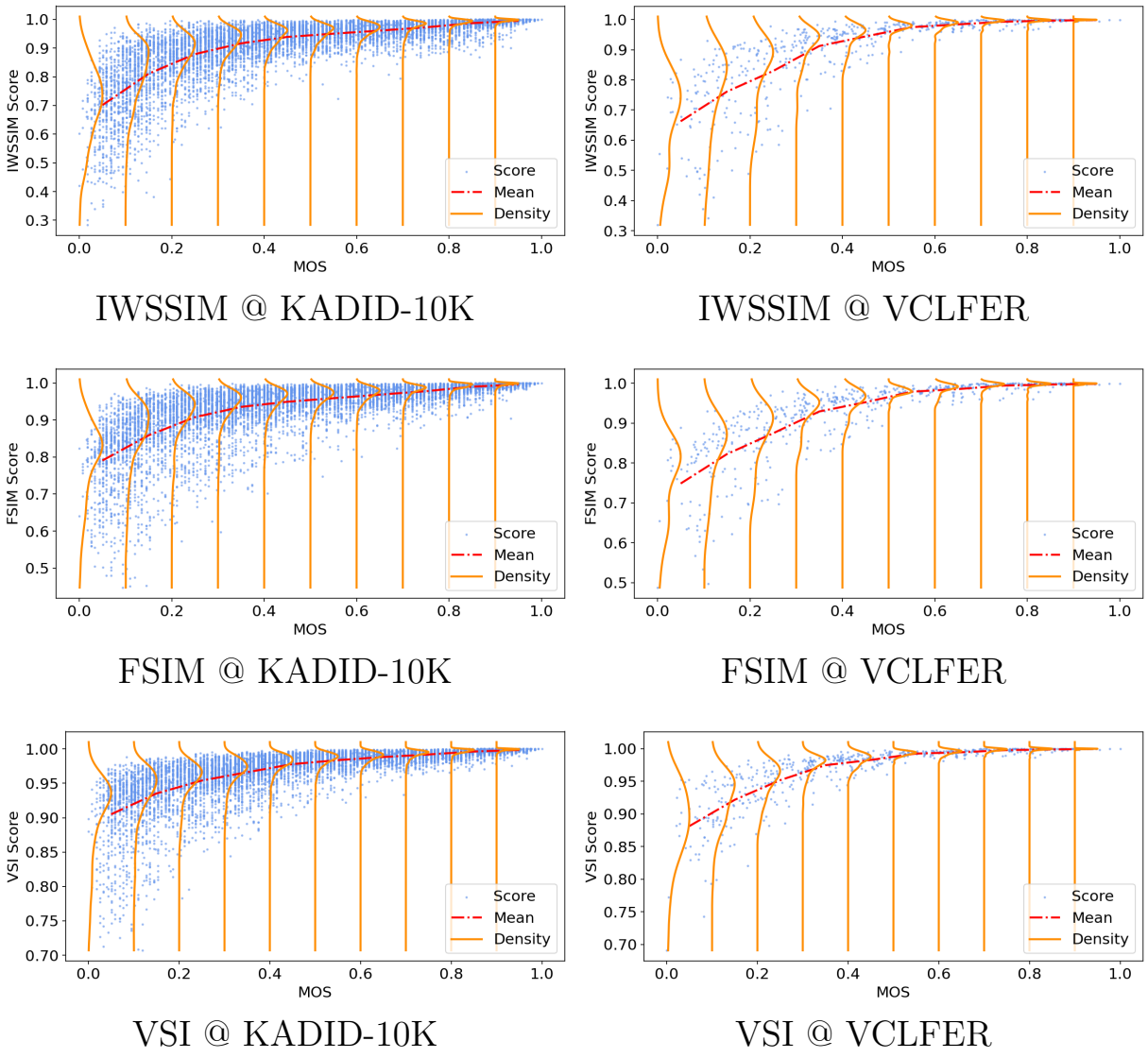
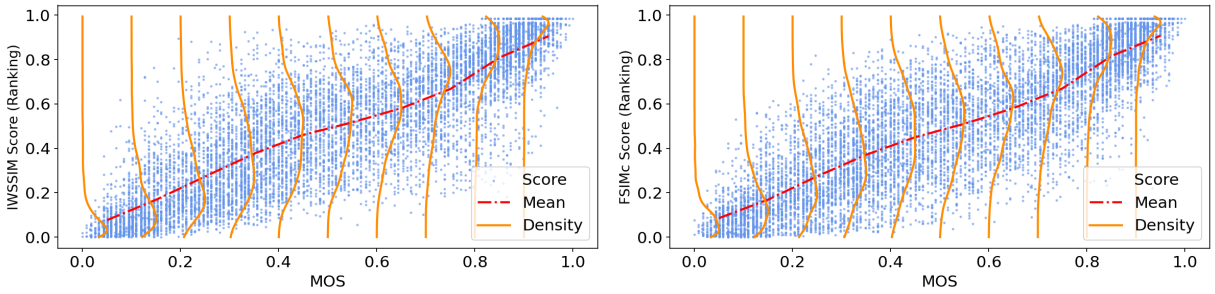


Figure 4.2: The empirical distribution of IWSSIM [4], FSIM [5] and VSI [6] evaluated on the KADID-10K [7] and VCLFER [8] dataset. The dashed red curve is the mean value of the scores (shown as blue dots). The solid orange curves are the conditional distributions, each representing the density of scores given a MOS range. It is easy to find that the conditional distributions are skewed toward the higher score side. Also, the variance of the conditional distribution is related to the MOS.



IWSSIM Rank @ KADID-10K FSIM Rank @ KADID-10K

Figure 4.3: The empirical distribution of the normalized rankings of IWSSIM [4] and FSIM [5] evaluated on the KADID-10K [7] dataset.

Table 4.1: Evaluation results in terms of SRCC of the proposed model, five other fusion methods and 16 individual IQA models used in the fusion. The average SRCC values are computed as weighted sums of individual ones, with weights determined by the number of images in each set. The top three best-performing models are shown in **bold** font while supervised methods are excluded.

Type	Model	LIVE r2	TID 2013	CSIQ	VCL @FER	CIDIQ 50	CIDIQ 100	MDID	MDID 2013	LIVE MD	MDIVL	avg SRCC
Individual FR-IQA	VSI[6]	0.9524	**0.8965**	0.9422	0.9317	0.7213	0.8106	0.8569	0.5700	0.8414	0.8269	0.8631
	FSIMc[179]	0.9645	0.8510	0.9309	0.9323	0.7608	0.8285	0.8904	0.5806	0.8666	0.8613	0.8628
	FSIM[5]	0.9634	0.8015	0.9242	0.9178	0.7438	0.8149	0.8872	0.5817	0.8635	0.8585	0.8628
	IWSSIM[4]	0.9567	0.7779	0.9212	0.9163	**0.8484**	**0.8564**	0.8911	**0.8551**	**0.8836**	0.8588	0.8559
	SFF[180]	0.9649	0.8513	**0.9627**	0.7738	0.7834	0.7689	0.8396	0.8005	0.8700	0.8535	0.8527
	DSS[181]	0.9616	0.7921	0.9555	0.9272	0.7755	0.8246	0.8658	**0.8078**	0.8714	**0.8759**	0.8520
	QASD[182]	0.9629	0.8674	0.9530	0.9231	0.7307	0.8079	0.7778	0.6687	0.8766	0.8315	0.8482
	MCSD[183]	0.9668	0.8089	**0.9592**	0.9224	0.7562	0.7808	0.8451	0.8269	0.8517	0.8370	0.8464
	CID MS[184]	0.9103	0.8314	0.8789	0.9366	0.8350	0.8062	0.8330	0.6168	0.8608	**0.8778**	0.8445
	GMSD[185]	0.9603	0.8044	0.9570	0.9177	0.7427	0.7675	0.8613	0.8283	0.8448	0.8210	0.8433
	VIF[47]	0.9636	0.6769	0.9194	0.8866	0.7203	0.6257	**0.9306**	**0.8444**	0.8823	0.8381	0.8024
	VIF DWT [186]	0.9681	0.6439	0.9020	0.8930	0.7224	0.5826	0.8943	0.7553	0.8479	0.8243	0.7768
Fused NR-IQA	HOSA[187]	**0.9990**	0.4705	0.5925	0.8574	0.4494	0.3248	0.6412	0.2993	0.6393	0.7399	0.5851
	NIQE[50]	0.9073	0.3132	0.6271	0.8126	0.3458	0.2212	0.6523	0.5451	0.7738	0.5713	0.5181
	MEON[188]	0.9409	0.3750	0.7248	0.9215	0.4101	0.2497	0.4861	0.2980	0.1917	0.5466	0.4969
	QAC[189]	0.8683	0.3722	0.4900	0.7686	0.3196	0.1944	0.3239	0.2272	0.3579	0.5524	0.4292
Empirical Fusion	CISI[127]	0.9680	0.8150	0.9425	0.9270	0.8231	0.8063	**0.9135**	0.6920	0.8740	0.8612	0.8634
	HFSIMc[126]	0.9610	0.8228	0.9423	0.9205	0.7315	0.7982	0.8202	0.5075	0.8624	0.8453	0.8345
	CM3[128]	0.9207	0.7136	0.8073	**0.9450**	0.6452	0.7659	0.7114	0.5055	**0.9206**	0.7733	0.7575
Rank Fusion	RRF[144, 143]	**0.9736**	0.6571	0.8421	0.9278	0.7873	0.7645	0.4941	0.3722	0.7022	0.8678	0.7133
	RRFW[145]	0.9648	0.7864	0.9245	0.9080	0.8055	0.8296	0.6153	0.4240	0.7352	0.8606	0.7874
Supervised	MMF [134]	0.9373	0.7475	0.7766	0.8482	0.2362	0.2180	0.7138	0.6275	-0.2498	-0.2741	0.5622
	CNNM [142]	0.9732	0.8737	0.9545	0.9125	0.7830	0.8008	0.9039	0.8125	0.8845	0.8433	0.8806
Proposed	**SF-ms**	0.9659	**0.8678**	0.9467	**0.9384**	**0.8473**	**0.8432**	0.9027	0.7963	**0.8895**	0.8553	**0.8869**
	RF-ms	0.9714	**0.8684**	**0.9619**	**0.9402**	**0.8443**	**0.8584**	0.9059	0.7484	0.8779	**0.8732**	**0.8897**
	SF-m	**0.9723**	0.8376	0.9506	0.9331	0.8379	0.8065	**0.9072**	0.8150	0.8799	0.8624	**0.8763**

Table 4.2: Evaluation results in terms of PLCC of the proposed model, five other fusion methods and 16 individual IQA models used in the fusion. The average PLCC values are computed as weighted sums of individual ones, with weights determined by the number of images in each set. The top three best-performing models are shown in **bold** font while supervised methods are excluded.

Type	Model	LIVE r2	TID 2013	CSIQ	VCL @FER	CIDIQ 50	CIDIQ 100	MDID	MDID 2013	LIVE MD	MDIVL	avg PLCC
Individual FR-IQA	VSI[6]	0.9428	**0.9000**	0.9279	0.9320	0.7226	0.8240	0.8703	0.5512	0.8789	0.8749	0.8707
	FSIMc[179]	0.9613	0.8769	0.9191	0.9329	0.7583	0.8410	0.8998	0.6412	0.8965	**0.9039**	0.8785
	FSIM[5]	0.9597	0.8589	0.9120	0.9185	0.7410	0.8265	0.8969	0.6474	0.8933	0.9037	0.8687
	IWSSIM[4]	0.9522	0.8319	0.9144	0.9191	**0.8476**	**0.8698**	0.8983	**0.8513**	**0.9109**	**0.9056**	0.8787
	SFF[180]	0.9632	0.8706	**0.9643**	0.7761	0.7834	0.7721	0.8590	0.7952	0.8893	0.8904	0.8658
	DSS[181]	0.9618	0.8530	**0.9612**	0.9259	0.7715	0.8267	0.8733	0.8168	0.9023	0.8973	0.8757
	QASD[182]	0.9574	0.8897	0.9481	0.9253	0.7257	0.8116	0.8063	0.6312	0.8966	0.8827	0.8625
	MCSD[183]	0.9675	0.8648	0.9560	0.9217	0.7532	0.7727	0.8637	0.8275	0.8847	0.8787	0.8705
	CID MS[184]	0.9159	0.8362	0.8732	0.9375	0.8364	0.8171	0.8414	0.6183	0.8917	0.8961	0.8511
	GMSD[185]	0.9603	0.8590	0.9541	0.9176	0.7387	0.7585	0.8776	**0.8309**	0.8808	0.8685	0.8671
	VIF[47]	0.9604	0.7720	0.9278	0.8938	0.7267	0.6415	**0.9367**	**0.8376**	0.9030	0.8736	0.8388
	VIF DWT [186]	0.9657	0.7657	0.9123	0.8969	0.7259	0.5845	0.9031	0.7531	0.8839	0.8653	0.8220
Fused NR-IQA	HOSA[187]	**0.9991**	0.5481	0.7240	0.8496	0.4969	0.3761	0.6521	0.2513	0.6768	0.7167	0.6275
	NIQE[50]	0.9052	0.4001	0.7170	0.8040	0.3703	0.2708	0.6728	0.5571	0.8387	0.5688	0.5642
	MEON[188]	0.9389	0.4946	0.7804	0.9221	0.4306	0.3854	0.5168	0.2430	0.2339	0.5722	0.5570
	QAC[189]	0.8625	0.4371	0.7067	0.7615	0.3573	0.2856	0.6043	0.4240	0.4145	0.5713	0.5338
Empirical Fusion	CISI[127]	0.9625	0.8575	0.9364	0.9289	0.8239	0.8193	**0.9220**	0.7095	0.9032	0.9007	0.8831
	HFSIMc[126]	0.9579	0.8635	0.9304	0.9211	0.7365	0.8120	0.8357	0.5242	0.8918	0.8893	0.8550
	CM3[128]	0.8337	0.6058	0.6870	**0.9435**	0.6383	0.7238	0.6362	0.4604	0.8718	0.8062	0.6893
Rank Fusion	RRF[144, 143]	0.9589	0.7393	0.8487	0.9269	0.7890	0.7752	0.6295	0.2440	0.7360	0.8958	0.7608
	RRFW[145]	0.9394	0.7722	0.8704	0.9148	0.7970	0.8214	0.5552	0.3252	0.7288	0.8786	0.7633
Supervised	MMF[134]	0.9248	0.7784	0.8751	0.8662	0.3145	0.2839	0.7250	0.6323	0.5451	0.3449	0.6776
	CNNM [142]	0.9712	0.8949	0.9648	0.9193	0.7871	0.8110	0.9120	0.8087	0.8913	0.8382	0.8904
Proposed	**SF-ms**	0.9635	**0.8919**	0.9448	0.9376	**0.8489**	**0.8734**	0.9073	0.7887	**0.9153**	0.8828	**0.9001**
	RF-ms	**0.9692**	**0.8921**	**0.9607**	**0.9394**	**0.8375**	**0.8696**	0.9117	0.7619	0.9067	**0.9062**	**0.9023**
	SF-m	**0.9723**	0.8631	0.9456	**0.9392**	0.8263	0.7910	**0.9232**	0.7830	**0.9126**	0.8758	**0.8863**

Type		IWSSIM	rrf_lt	rrf_w_lt	MMF	CNNM	SF-ms	RF-ms	SF-m
Single Metric	IWSSIM	—	011-11111-	1111111111	1111111111	0000-101–	0000–01-1	0000–01–	0000-101-1
Rank	rrf_lt	100-00000-	—	1001-01–1	1001110011	000000000-	000000000-	000000000-	00000-0001
Fusion	rrf_w_lt	0000000000	0110-10–0	—	0–1110011	00000-0000	000000000-	0000000000	000001000-
Supervised	MMF	0000000000	0110001100	1–0001100	—	0000000000	0000000000	0000000000	0000000000
Fusion	CNNM	1111-010–	111111111-	11111-1111	1111111111	—	101–01–1	101–01—	-11-11—1
	SF-ms	1111–10-0	111111111-	111111111-	1111111111	010–10–0	—	0-0——0	01–110—
Proposed	RF-ms	1111–10–	111111111-	1111111111	1111111111	010–10—	1-1——1	—	-11–10–1
	SF-m	1111-010-0	11111-1110	111110111-	1111111111	-00-00—0	10–001—	-00–01–0	—

Table 4.3: Statistical significance testing of fusion methods using prediction residuals. Each entry is a codeword made up of ten symbols, with each symbol representing the test result for an IQA database. The order of the database is the same as in Table 4.1. 1 means that the method is statistically better than the method in the column on that particular dataset, 0 means that it is statistically worse, and - means that it is statistically indistinguishable.

Table 4.4: Evaluation results when "bad" IQA metrics are added to the list of models to be fused.

Type	Model	avg SRCC	avg PLCC
Rank Fusion	RRF[144, 143]	0.6096	0.6467
	RRFW[145]	0.7171	0.7353
Proposed	**SF-ms**	**0.8859**	**0.9003**
	RF-ms	**0.8842**	**0.8987**
	SF-m	**0.8745**	**0.8928**

Table 4.5: Evaluation results when IQA datasets with "bad" MOS are used as the training set.

Type	Model	avg SRCC	avg PLCC
Supervised	MMF[134]	0.5477	0.6478
	CNNM [142]	0.8409	0.8624
Proposed	**SF-ms**	**0.8869**	**0.9001**
	RF-ms	**0.8897**	**0.9023**
	SF-m	**0.8763**	**0.8863**

Table 4.6: Evaluation results when small IQA datasets are used as the training set.

Type	Model	avg SRCC	avg PLCC
Supervised	MMF[134]	0.4703	0.6067
	CNNM [142]	0.7988	0.8118
Proposed	**SF-ms**	**0.8827**	**0.8978**
	RF-ms	**0.8880**	**0.9018**
	SF-m	**0.8588**	**0.8783**

Table 4.7: Evaluation results of FQA score fusion. We fused seven FQA methods on the FocusPath [2] dataset. The top three best-performing models are shown in **bold** font.

Type	Model	SRCC	PLCC
Individual FQA	FQPath [57]	0.8384	0.8268
	HVS-MaxPol-1 [2]	0.8035	0.8003
	HVS-MaxPol-2 [2]	0.8421	0.8310
	Synthetic-MaxPol [58]	0.8216	0.8111
	LPC [107]	0.8298	0.8259
	GPC [108]	0.7681	0.7430
	SPARISH [116]	0.3195	0.3386
Rank Fusion	RRF[144]	0.8376	0.8241
	RRFW[145]	0.8376	0.8085
Supervised	MMF[134]	0.8075	0.8016
	CNNM [142]	**0.8761**	**0.8713**
Proposed	**SF-ms**	**0.8595**	**0.8478**
	RF-ms	**0.8625**	**0.8521**
	SF-m	0.8467	0.8355
	RF-m	0.8524	0.8403

Table 4.8: Inference speed comparison of different fusion methods. The time reported is for fusing a single image based on evaluations from 16 IQA models. Note that the inference speed of rank fusion methods is also influenced by the number of images in the dataset, which is set to 10,000 in this experiment.

Type	**Model**	**Speed (s)**
Rank Fusion	RRF[144]	1.17e−2
	RRFW[145]	12.26
Supervised	MMF[134]	1.11e−4
	CNNM [142]	2.04e−6
Proposed	MLP Encoder	2.08e−6
	Cascade Neural Network Encoder	2.04e−6
	Linear Encoder	1.16e−8

Chapter 5

Whole Slide Image Virtual Refocusing

5.1 Introduction

Pathology is the study and diagnosis of disease, which involves the examination of surgically removed organs, tissues, or bodily fluids. Subfields of pathology include histology and cytopathology etc. Histology examines surgically removed tissues under a microscope. Cytopathology studies diseases on the cellular level, which involves the examination of free cells from bodily tissues or fluids. Both histology and cytopathology require the specimen to be processed to satisfy the requirements of microscopic viewing. No matter what kind of pathology, a professional pathologist must be present to conduct the examination. To enhance the efficiency of the diagnosis workflow, digital pathology emerged during the 1960s. Whole slide imaging is a core step in digital pathology, where traditional glass slides are digitized into high-resolution images that can be viewed, stored, shared, and analyzed on computer systems. Whole slide imaging revolutionizes digital pathology, facilitating remote diagnosis and collaborative research. However, ensuring high-quality scans is a challenging task due to defocus.

In WSI scanners, bright-field microscopes are commonly employed for object magnifi-

cation. The resolving power of such microscopes is quantifiable through optical resolution (R), defined as the shortest distance between two Airy disks discernible on the image plane (Eq. 3.2). High resolution (a low R value), however, is often accompanied by a trade-off: reduced DOF. DOF represents the range surrounding the focal plane within which image sharpness remains consistent [164]. Objects that lie within the DOF will appear sharp, whereas those outside this range will be out of focus. The relationship between R and DOF can be expressed as

$$DOF = \frac{2R}{tan(\alpha)} \tag{5.1}$$

where α is the half angle of the cone of light entering the objective lens [164]. It is easy to find that the DOF is proportional to the resolution R, meaning most microscopes will have a very shallow DOF. Take the most commonly used 40X objective lens with a 0.65 NA as an example, the DOF is around $1\mu m$. The major disadvantage of using a shallow DOF lens is that it is challenging to capture an all-in-focus image, meaning that the image is in-focus at every pixel.

Depending on the focus plane setting, tissue surface evenness, and tissue thickness, three out-of-focus scenarios can arise:

- First, an incorrect focus plane setting can lead to an entirely out-of-focus image. This typically arises from autofocus system malfunctions. Physical distortions, such as bubbles, marker ink or dust particles on the coverslip, can also mislead the autofocus system.
- Second, even with a correct focus plane setting (positioned directly beneath the coverslip's bottom surface), non-uniform tissue surface, tissue folding, or the presence of bubbles can cause certain tissue regions to fall outside the DOF. Consequently, the resulting image will be partially out of focus.
- Third, even when the focus plane is correctly set and the tissue surface is even, the shallow DOF might not cover the entire tissue thickness, which is $4\mu m$ to $5\mu m$ for

[1]The objective lens icon is adopted from Biorender.

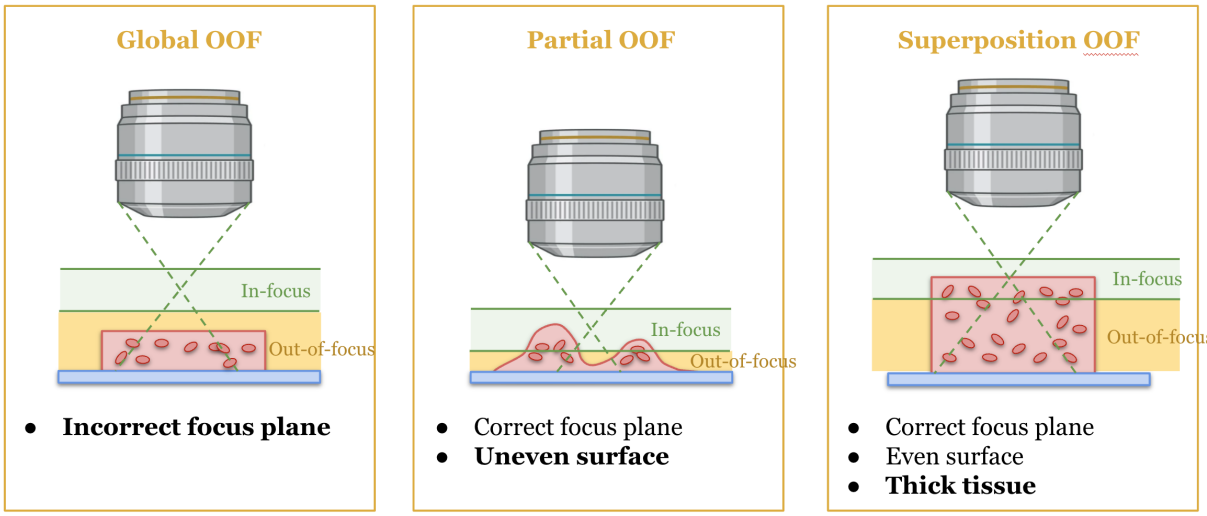


Figure 5.1: Illustration of three types of Out-of-Focus. The green area is the DOF. The tissue that lies within it will be in-focus, while those that lie outside of it will be out-of-focus [1].

most paraffin fixed ones [199]. Due to the complete illumination of the tissue by the light source, out-of-focus light traversing deeper tissue layers outside the DOF interferes with the in-focus light. The captured image, therefore, becomes a superposition of in-focus and out-of-focus components, leading to blurriness. Therefore, we name this type of out-of-focus as superposition out-of-focus.

In traditional microscopy, the focus problem might be solved by manually adjusting the focus while the pathologist inspects the slide. In digital pathology, this problem gets more challenging. For the first two out-of-focus scenarios, z-stack imaging offers a potential solution. Z-stack (also referred to as focal stack) is a technique that captures a series of images at different focal planes (also referred to as z-level or focus level). Z-stack increases the probability of capturing the in-focus image within one or more z-levels. However, the scanning time and data volume increase proportionally to the number of z-levels. However, the trade-off is a proportional increase in scanning time and data volume. Considering that a single z-level WSI scanning is already very time-consuming and consumes a lot of storage, scanning multiple z-levels will worsen the situation. Consequently, z-stack is

Type	Global OOF	Partial OOF	Superposition OOF	WSI	Note
Confocal	Yes	Yes	Yes	No	Fluorescence only
Light Sheet	Yes	Yes	Yes	No	Fluorescence only
Light Field	Yes	Yes	Maybe	No	Low axial resolution
Bright Field (Refocus)	Yes	Yes	Maybe	Yes	No extra equipment

Table 5.1: Comparison of different types of microscopes and their abilities to handle the three types of out-of-focus. However, most methods are not directly applicable to WSIs in digital pathology due to practical constraints and the need for specialized, often expensive, equipment.

primarily employed for research purposes or specialized organ imaging. Nevertheless, when scanning at a lower number of z-levels, or when the distance between adjacent z-levels gets larger, it is still not guaranteed that z-stack can always capture an in-focus image.

The third scenario, characterized by blurriness from the superposition of in-focus and out-of-focus light, necessitates an understanding of the tissue's 3D structure. Traditional 2D imaging proves inadequate in this context. 3D imaging techniques such as confocal microscopy, light sheet microscopy, two-photon microscopy, structured illumination microscopy can be used to capture the 3D information. Light field microscopy captures partial 3D information and can reconstruct the 3D structure with deconvolution. However, these microscopes are typically more expensive and exhibit limitations such as reduced image capture speed or constraints in lateral/axial resolution. Furthermore, some of these techniques are restricted to fluorescent samples. Therefore, 2D bright-field microscopy, despite its inherent limitations, remains the dominant technique employed by WSI scanners in digital pathology. Although it can not solve the superposition out-of-focus issue, one can still observe the 3D biostructure by adjusting the focus plane. Table 5.1 provides a detailed comparison of various microscopy techniques.

Image Deblurring/Deconvolution Image deblurring/deconvolution presents a practical solution to the global out-of-focus scenario. Traditional image deblurring methods [153] employ iterative techniques to estimate the blur kernel and subsequently restore the

image. Recent advances leverage deep learning for WSI deblurring, restoring in-focus images in an end-to-end fashion [154, 155, 156, 157, 200, 201]. These models are trained on paired datasets of out-of-focus and in-focus images. This approach is effective when the tissue surface is even and the defocus level is consistent throughout the image, a characteristic of the global out-of-focus scenario. In such cases, acquiring training image pairs is straightforward. However, when defocus levels vary across the image, in-focus regions are distributed across multiple z-levels. This represents the partial out-of-focus case. Consequently, capturing a ground truth image with every pixel in focus using standard bright-field microscopes becomes challenging. Some studies attempt to circumvent this issue by performing deblurring at the patch level, assuming uniform defocus within each small patch. However, this strategy often leads to inconsistencies between patches, as global image context and consistency are disregarded. Patch-based deblurring is also susceptible to distortions such as checkerboard artifacts. A more comprehensive review of WSI deblurring methodologies is presented in Section 2.3.1. Nevertheless, both image deblurring or deconvolution aim to generate an AIF image, which collapses the desired 3D tissue structure.

Focus Interpolation Focus interpolation [161, 162, 163] offers an alternative to deblurring, addressing both global and partial out-of-focus scenarios. This technique allows for the synthesis of the intermediate z-level from two captured z-levels. Through iterative interpolation, continuous focus adjustment becomes feasible, a capability typically absent in conventional z-stack imaging. Furthermore, focus interpolation significantly reduces scanning time and storage requirements, as intermediate z-levels can be generated on demand. Section 2.3.2 provides a more in-depth review of focus interpolation methods. However, despite its advantages, focus interpolation suffers from several limitations:

- It is inherently restricted to generating the middle z-level between two input z-levels. Synthesizing arbitrary intermediate z-levels necessitates iterative interpolation, a process that is computationally expensive and prone to error accumulation, leading to substantial prediction errors.

- Focus interpolation lacks extrapolation capabilities. The target z-level must lie within the range of the input z-levels. In practice, this constraint is difficult to satisfy, as

Modality	Task	Objective	Input(s)	Transparent Object	Continuous Refocus	Pixel-wise Refocus
WSI	WSI Virtual Refocus	Continuous Focal Stack	Sampled Images in Focal Stack	Yes	Yes	Yes
	WSI Focus Interpolation	Middle Focus Image	Two Images in Focal Stack	Yes	No	No
Natural Image	Natural Image Refocus	Continuous Focal Stack	AIF image + Depth Map	No	Yes	Yes
	Multi-focus Fusion	AIF Image	Focal stack + Depth Map	No	N.A.	N.A.
	All-in-Focus Image	AIF Image	Focal stack + Depth Map	No	N.A.	N.A.
	2D Deconvolution	AIF Image	An Out-of-Focus Image	No	N.A.	N.A.

Table 5.2: Comparison of different tasks related to WSI virtual refocusing. Natural image-based tasks assume scene non-transparency, making them unsuitable for WSI. The goal in WSI is to visualize 3D tissue structure. Generating a single AIF image, however, collapses this 3D information, hiding the desired structural details.

the precise z-level of the in-focus image is often unknown.

- When multiple input images are available, focus interpolation can only utilize two at a time. While the z-level pair closest to the target provides the most relevant information, the remaining images may still contain valuable data that is disregarded.
- Focus interpolation is inapplicable when only a single input image is available.

Virtual Refocusing Physical refocusing involves adjusting the focal plane using 3D information acquired during image capture. In microscopy techniques that capture partial 3D data, such as light field microscopy, physical refocusing can be achieved by directly rendering the desired z-level. Certain 3D microscopy methods, including confocal microscopy, light sheet microscopy, and two-photon microscopy, enable optical sectioning, allowing the direct acquisition of 2D radiance at arbitrary z-levels. In these cases, physical refocusing becomes unnecessary. However, as described before, most of these 3D microscopy methods are not applicable to WSI used in digital pathology.

Virtual or digital refocusing aims to approximate the focus adjustment process without relying on an explicit 3D model. In contrast to focus interpolation, it requires only a single input image. This makes it more flexible and practical to use. Refocusing methods for natural images rely on an AIF image and a depth map that describes the distance of the scene from the image plane. By synthesizing blur on an AIF image according to the

depth map, one can achieve continuous focus adjustment. This essentially assumes that the scene is non-transparent. What it is refocusing is just the 2D surface of the 3D scene. This limitation makes it not suitable for WSI refocusing.

In handle refocusing in transparent scenes, we propose the first virtual refocusing method for WSI. Its input is an arbitrary number of images in a focal stack and a target focus plane/map. Tasks such as WSI deblurring, WSI out-of-focus synthesis, WSI focus interpolation, WSI multi-focus fusion and WSI AIF generation can be considered as special cases of the proposed virtual refocusing framework. A detailed comparison of the related tasks is provided in Table 5.2. More importantly, it simulates the way pathologists inspect slides, adjusting the focus level to accommodate uneven tissue surfaces and thick tissues. Moreover, virtual refocusing exhibits greater flexibility than physical refocusing, as it can refocus each pixel individually to a target focus plane.

In fluorescence microscopy, Deep-Z [202] has been applied to refocus a 2D image to a user-defined focus plane. The appearance of fluorescence images differs from WSIs in two major aspects: 1) Fluorescence images are sparser where more pixels are black compared to WSI. This is because the fluorescent stains only bind to a specific molecule in the tissue. As a result, the occlusion between different focus layers is less prominent in fluorescence images. 2) Fluorescence microscopy relies on detecting light emitted from fluorescent stains, where the signal is usually weaker than bright field microscopy. This results in low SNR images that are less detailed. Nevertheless, WSI virtual refocusing has not been studied in the literature. Compared to fluorescence microscopy, WSIs have more delicate structures and generally contain more information. Since the occlusion phenomenon is more noticeable in WSI, a deeper understanding of the 3D structure is needed for refocusing, which is missing in the fluorescence case [202]. In terms of model design, it omits the image formation model and does not have a 3D radiance field reconstruction step. It also accepts only one input image, which does not capture the rich 3D information within the z-stack. It is also limited to a uniform target focus plane, which is less practical in situations where the tissue landscape is uneven.

Complex 3D Image Formation Model Although virtual refocusing seems to be an all-in-one solution to the out-of-focus restoration and defocus synthesis problem, as far as we know, there is barely any refocus model for WSI. Most refocus models are designed

for natural images, which fundamentally differ from WSI: the objects in most natural scenes have reflective surfaces. However, tissue slides are translucent or transparent. For the reflective scene, the captured image can be described using a depth map and an AIF image. By convolving the AIF image with proper PSFs that correspond to the depth, we can generate images captured in other focal planes [203, 204, 205]. This means rather than the actual 3D model of the scene, we only need to consider the 2D manifold represented by the depth map and the AIF image. On the other hand, since tissue slides are translucent or transparent, each captured image is a superposition of infinite AIF images at all focal planes convolved with corresponding PSFs. This means that in order to refocus a WSI exactly, the real 3D model of the tissue and related PSFs are needed.

Besides the complex 3D model required for exact refocusing of WSIs, the PSFs involved also differ from natural images. In natural image refocusing, once the lens specifications are determined, the PSF is approximately a function of focus depth alone. The image formation process can be formulated as

$$I(x,y)|_{z_0} = R(x,y) * h(z_0 - d(x,y)) \tag{5.2}$$

where $I_{z_0} \in \mathbb{R}^{H\times W\times 3}$ is the final captured image when setting the focus plane z at z_0. Let $R \in \mathbb{R}^{H\times W\times 3}$ denote the 2D radiance field, often regarded as the AIF image. The symbol $*$ represents the 2D convolution operator. While factors such as atmospheric interference (e.g., fog, haze, rain) can influence the precise PSF, they are typically negligible. The relative simplicity of the PSF in natural image refocusing contributes to its ease of implementation. However, in tissue slides, the PSF can be considered a function of both depth and the 3D tissue structure. For a given focus plane, the planes above and below can be conceptualized as imperfect lenses composed of the FFPE tissue. The tissue's shape, size, and spatially varying refractive, reflective, and absorption coefficients all contribute to the final radiance field. Modeling the optical characteristics of each tissue layer is a considerable challenge due to the inherent complexity and heterogeneity of biological samples. Even with a complete 3D model, rendering the final image remains a computationally intensive process that often requires ray tracing techniques rather than simple convolution [206, 207]. Assuming an incoherent light source, the image formation process

in transmission microscopy can be expressed as:

$$I(x,y)|_{z_0} = \int_0^D R(x,y,z) * h(z_0 - z)dz \tag{5.3}$$

where $R \in \mathbb{R}^{H\times W\times D\times 3}$ represents the 3D radiance field resulting from the interaction of the light source with the physical 3D tissue model. The term h denotes the 3D PSF of the microscope lens, and $h(z_0 - z)$ is an axial slice of this PSF. D represents the maximum tissue thickness. The complexities inherent in modeling 3D tissue structures and rendering the corresponding 3D radiance field make WSI refocusing a considerably more complex task compared to natural image refocusing. A limited number of studies [208, 209] have explored 3D reconstruction of biological specimens from z-stacks using a MAP framework. However, these studies employed a total variation prior, which provides limited information and inadequately captures the true 3D tissue distribution. Furthermore, [209] neglected refraction and reflection effects in computing the 3D radiance field, while [208] bypassed 3D radiance field computation altogether, substituting it with the 3D model. Consequently, these methods are only effective for relatively simple, synthetic datasets lacking complex 3D structures and optical characteristics. WSI refocusing via 3D model reconstruction and 3D radiance rendering is not only computationally demanding but also inherently challenging. We introduce a novel, end-to-end WSI refocusing model that circumvents the need for complex 3D modeling and radiance field rendering.

Implicit 3D Radiance Field Reconstruction As demonstrated by the image formation process in Eq. 5.3, each captured image arises from the convolution of the 3D tissue radiance field with a set of 3D PSFs at different axial positions. Consequently, accurate estimation of the focus stack requires both the 3D radiance field and the PSFs. However, capturing the 3D radiance field is inherently challenging without specialized tissue processing techniques and imaging equipment. Additionally, acquiring the PSFs for each objective lens at all axial levels is a laborious process. To address these challenges, we introduce a novel 3D Radiance Consistency Loss (RCLoss) that guides the model to implicitly learn the 3D tissue representation. This approach eliminates the need for any additional information, such as explicit 3D model or PSFs, relying solely on the focal stack images. The projection of the 3D radiance field onto the image plane results in substantial

information loss. Consequently, a single image at any focus level cannot fully represent the 3D radiance field. A logical extension is to enable the model to accept multiple images from a focal stack as input. These images, being projections of the same 3D radiance field convolved with different PSF sets, contain rich information that can enhance the model's 3D representation learning. To ensure model versatility, the number of input images should not be fixed. Our proposed model accepts a variable number of input images, ranging from 1 to 16, achieving a balance between representation capacity and computational efficiency.

Conditioned Refocusing Using the 3D Representation Having established a rich 3D representation, the next step is to project and collapse this 3D feature into a 2D image, guided by a focus map. To achieve this, we introduce a refocus module incorporating a novel Focal Stack Cross-Attention Pooling (FSCA Pool) mechanism. Reducing a 3D feature to 2D necessitates a pooling operation along the depth dimension. Traditional max pooling or mean pooling disregards the inherent 3D structure within the representation, contradicting our goal of reconstructing the 3D radiance field. The FSCA Pool module employs a cross-attention mechanism to selectively extract the most relevant information based on the target focus map. This approach aligns with the inherent structure of the 3D representation, seamlessly integrating with the 3D radiance field reconstruction module.

WSI Distance Measure Traditional image distance or FR IQA metrics, such as MSE, MAE, and SSIM [13], have been widely adopted as loss functions in various image reconstruction tasks. However, these metrics primarily focus on low-level visual features, neglecting semantic information crucial for IQA. Deep learning-based FR IQA models [11, 12] have emerged to address this limitation by incorporating both low-level and high-level features. These models typically achieve this by assessing differences in features extracted at multiple levels of a deep neural network. While these powerful loss functions have proven effective in WSI deblurring and interpolation tasks, the pre-trained feature extractors they employ are designed for general natural images. For instance, LPIPS [11] and DISTS [12] utilize a VGG16 network [210] pre-trained on the ImageNet dataset [211] for object recognition. However, ImageNet [211] solely comprises natural images, which differ significantly from WSIs in several key aspects:

- Objects of interest: Natural images depict everyday scenes and objects, while WSIs

focus on microscopic tissue structures.

- Imaging devices: Natural images are typically captured using standard cameras, whereas WSIs employ specialized microscopes.
- Illumination conditions: Natural images are subject to varying natural lighting, while WSIs are captured under controlled microscope illumination.
- Image post-processing: WSIs often undergo specific post-processing steps, such as color normalization and artifact removal, which are not typically applied to natural images.

Further justifications for these distinctions are provided in Section 3.2. Despite their success in natural image-related applications, these loss functions may struggle when applied to WSIs, as they lack knowledge of such data. To bridge this gap, we propose a novel image distance metric specifically tailored for WSIs. Our results demonstrate that incorporating this metric as the loss function significantly enhances model performance compared to its natural image counterparts.

In summary, the key contributions of this model are:

1. To the best of our knowledge, this is the first WSI virtual refocusing model.
2. The model implicitly learns a 3D radiance field representation through a novel 3D consistency constraint. It accepts an arbitrary number of focal stack images as input, enriching the 3D information learned while being practical.
3. A novel focal stack cross-attention module allows the model to selectively extract information relevant to the target focus map, facilitating the generation of refocused images.
4. The model enables continuous refocusing of individual pixels to different focal planes, which is more flexible than physical refocusing.

5. Tasks such as WSI deblurring, out-of-focus synthesis, focus interpolation, multi-focus fusion and AIF generation can be considered as special cases of the proposed framework. This is also the first WSI defocus synthesis model to incorporate realistic defocus distortions.

6. A novel, WSI-specific image distance metric, employed as the reconstruction loss, significantly enhances model performance.

5.2 Method

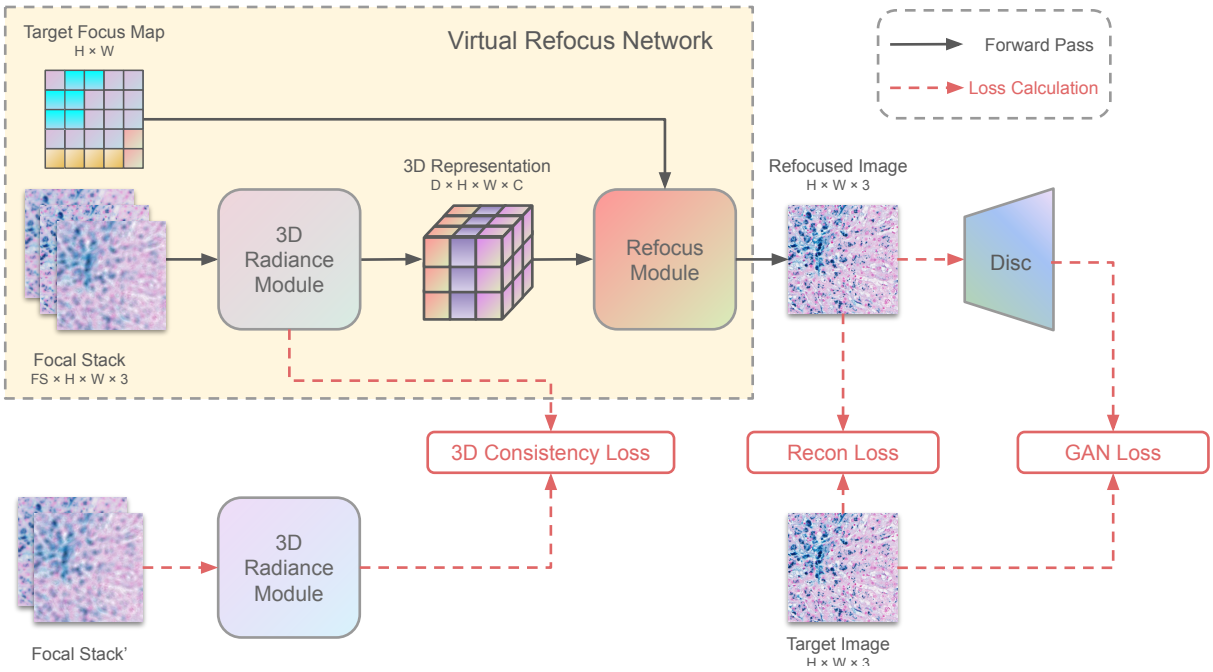


Figure 5.2: Overview of the proposed virtual refocusing network. The network consists of two major components: (1) a 3D radiance field module for implicitly learning a 3D tissue representation and (2) a refocus module that refocuses this 3D representation based on a 2D focus map.

5.2.1 General Overview

The proposed model adopts an encoder-decoder architecture. Inputs consist of a focal stack comprising $FS \in \mathbb{N}^+$ images, denoted as $X \in \mathbb{R}^{FS\times H\times W\times 3} = \{x_i \in \mathbb{R}^{H\times W\times 3} | i = 1, \cdots, FS\}$, their corresponding focus maps $F \in \mathbb{R}^{FS\times H\times W\times 1} = \{f_i \in \mathbb{R}^{H\times W\times 1} | i = 1, \cdots, FS\}$, and a target focus map $f_t \in \mathbb{R}^{H\times W\times 1}$. The focus maps ($F$) are optional. If not provided, they can be predicted using pre-trained FQA models such as FocusLiteNN [176]. The model's objective is to generate a refocused image $x_t \in \mathbb{R}^{H\times W\times 3}$ where the focus distance of each pixel is determined by the target focus map f_t. Figure 5.2 provides an overview of the proposed model. The model is composed of two primary modules: a 3D radiance field module and a refocus module. The 3D radiance field module implicitly learns a 3D representation of the tissue from multiple focal stack images. This process relies solely on the focal stack images, without requiring any additional information such as explicit 3D structural data or PSFs. Detailed descriptions of this module are presented in Section 5.2.2. The refocus module aims to refocus the 3D representation based on the guidance provided by the 2D focus map. Detailed descriptions of this module are provided in Section 5.2.3. An additional discriminator is incorporated to enhance the visual quality of the generated refocused image by adversarial learning.

The model is trained using a supervised learning approach. Input images and the corresponding target image are randomly sampled from the same focal stack. The loss function comprises three components: a reconstruction loss, a 3D consistency loss, and an adversarial training loss. The reconstruction loss is designed to minimize the perceptual difference between the refocused image and the target image. Further details are provided in Section 5.2.4. The 3D consistency loss guides the 3D radiance field module in learning a 3D representation of the tissue, as detailed in Section 5.2.2. The adversarial training loss encourages the model to generate more realistic images, with details outlined in Section 5.2.5.

The U-net encoder-decoder architecture [160] is a popular choice for image-to-image tasks such as image restoration [212] and style transfer [213]. U-net employs skip connections, directly transferring intermediate features from the encoder to the decoder, bypassing intermediate layers. This design offers two key advantages: (1) preserving spatial details

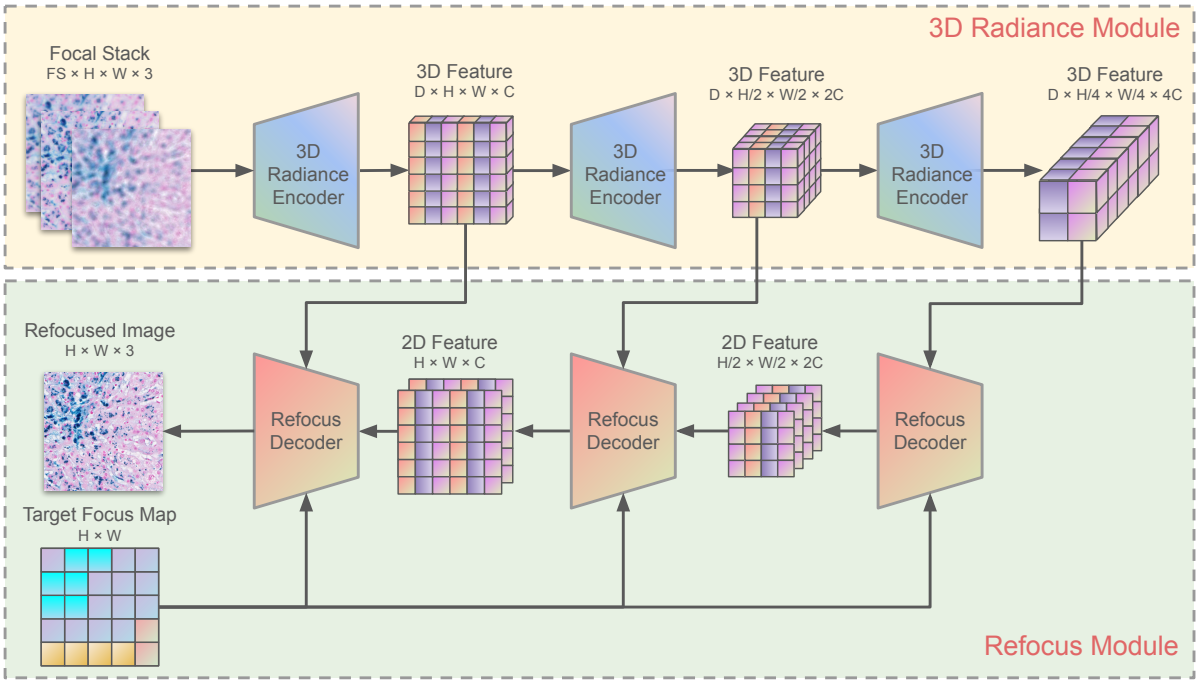


Figure 5.3: Architecture of the 3D radiance field module and refocus module. The overall model design is based on a U-net architecture. Both modules consist of three stages, with skip connections linking the outputs of each stage in the 3D radiance filed module to the corresponding stage in the refocus module.

and (2) enabling the training of deeper networks by mitigating the vanishing gradient problem. Our proposed model also utilizes a U-net architecture, as shown in Figure 5.3. The 3D radiance field module is composed of three 3D encoders. The first encoder maps the input images to a latent representation with dimensions $D \times H \times W \times C$, where D is the depth of the 3D representation and C is the number of channels. The subsequent two encoders further downsample this latent representation spatially, resulting in output features of size $D \times \frac{H}{2} \times \frac{W}{2} \times 2C$ and $D \times \frac{H}{4} \times \frac{W}{4} \times 4C$, respectively. Section 5.2.2 provides a detailed description of the encoder architecture. The refocus module consists of three decoders, each corresponding to an encoder in the 3D radiance filed module. Each decoder receives three inputs: the corresponding 3D representation from the 3D radiance field module, the 2D

features pooled by the preceding decoder (if available), and the target focus map. The first two decoders also spatially upsample the features to restore the original spatial resolution. Detailed descriptions of the decoder architecture can be found in Section 5.2.3.

Other than network design, the choice of loss function is critical in training image-to-image models. Commonly used loss functions, such as MSE, MAE, and SSIM [13], only consider low-level visual features, neglecting semantic information essential for capturing perceptual image similarity. Deep learning-based FR IQA or image distance metrics [11, 12] address this limitation by evaluating distances at both low-level and high-level feature representations. This is typically achieved by calculating a weighted sum of distances between features extracted at multiple levels of a deep neural network. These deep learning-based loss functions have found application in WSI deblurring and interpolation models. However, the pre-trained feature extractors used in these metrics are designed and trained exclusively on natural images. For example, both LPIPS [11] and DISTS [12] employ a VGG16 network [210] trained on the ImageNet dataset [211] for object recognition task. As discussed in Section 3.2, ImageNet [211] consists solely of natural images, which exhibit significant differences from WSIs. While effective for natural images, these loss functions may not be optimal for WSIs due to this domain mismatch. To mitigate this issue, we propose a novel WSI-oriented image distance metric based on DISTS [12]. Our experiments demonstrate that using this metric as the loss function significantly improves model performance compared to its natural image counterpart. A detailed description of this metric is presented in Section 5.2.4.

5.2.2 Implicit 3D Radiance Field Reconstruction

3D Radiance Field Encoder

According to the image formation process illustrated in Eq 5.3, we could synthesize a refocused image by convolving the 3D radiance field with the 3D PSFs corresponding to the target focus map. However, obtaining an accurate 3D radiance field and the PSFs are both very challenging. Recognizing that our ultimate goal is the refocused image, not the explicit 3D radiance field, we propose to learn a latent representation of the 3D radiance

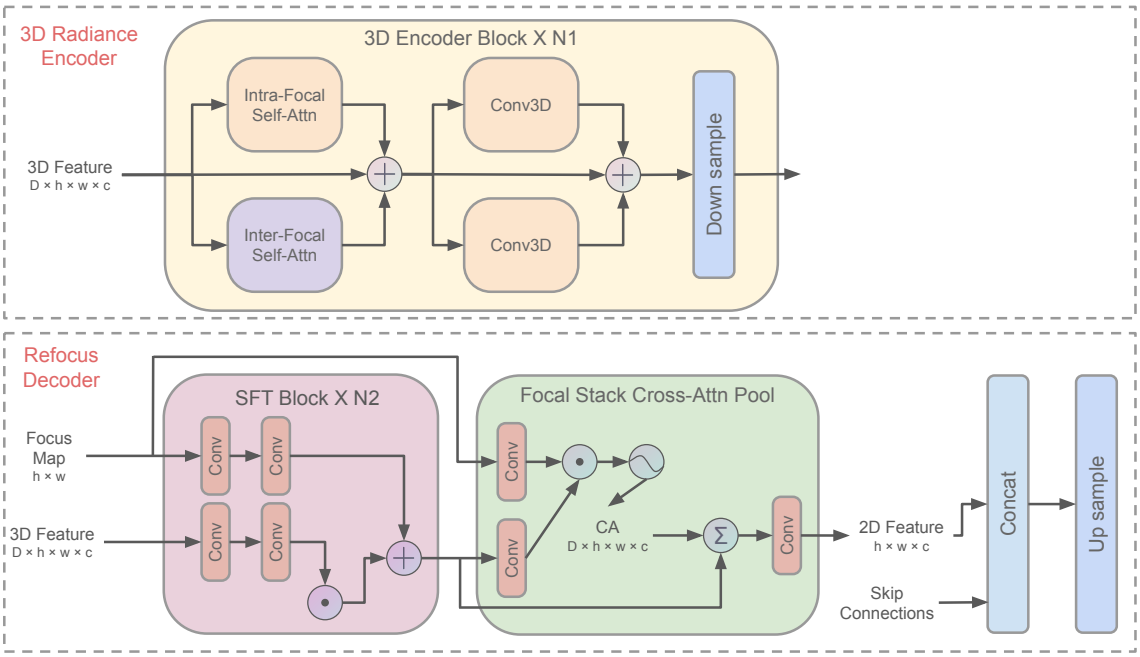


Figure 5.4: Network structure of the 3D radiance encoder block and refocus decoder block. Each encoder block incorporates intra-image and inter-image attention modules, followed by 3D convolutions. Each decoder block consists of an SFT block and a FACA Pooling layer.

field instead of directly modeling it. This is accomplished using a 3D radiance field module and a 3D radiance consistency loss.

The 3D radiance field module comprises three 3D radiance encoders. These encoders aim to capture inter-image and intra-image correlations within the focal stack. Considering that different images in a focal stack are generated by the same radiance field, these images share a lot of information about the underlying radiance. By leveraging these inter-relationships, we can extract 3D information from the features, potentially revealing the radiance field. Furthermore, the 3D radiance field features are structured in the depth dimension, meaning the top layers in this feature represent the surface of the tissue and vice versa. Consequently, it is intuitive to use 3D convolutions in this encoder.

Intra-image correlations, similar to self-attention mechanisms used in 2D images, explore spatial dependencies among pixels. In the context of out-of-focus images, the circle of confusion is a point light source projected onto the image plane. The size of the circle of confusion is determined by the defocus level. To capture this defocus characteristic, we need to assess intra-image correlations within a large receptive field. While 2D convolution can partially achieve this, its receptive field within a single layer is limited. Intra-image self-attention offers a broader receptive field by establishing pairwise correlations among pixels.

Figure 5.4 illustrates the architecture of the 3D radiance encoder. Each encoder consists of an intra-image self-attention module, an inter-image self-attention module, and two 3D convolution layers. To mitigate the computational cost of 3D convolutions and attention mechanisms, we adopt an efficient modification proposed in [214]. This modification decomposes 3D attention and 3D convolution into separable spatial-wise and depth-wise operations. By processing data along these two branches and subsequently fusing the results, computational complexity is significantly reduced while maintaining comparable performance.

3D Consistency Loss

As discussed previously, the network design of the 3D radiance field module facilitates the capture of 3D information and lens defocus characteristics. However, we require a more explicit mechanism to guide the learning of a 3D radiance field representation. To this end, we propose a novel 3D radiance consistency loss that leverages the inherent invariance within a focal stack. Consider two disjoint sets of images sampled from the same focal stack $X = \{x_i \in \mathbb{R}^{H\times W\times 3} | i = 1, \cdots, N\}$. Let these sets be denoted as $X_m \in X$ with $|X_m| = m$ and $X_n \in X$ with $|X_n| = n$, where $X_m \cap X_n = \Phi$. Given that images in the same focal stack are generated by convolving the same radiance field with different 3D PSFs, these two subsets naturally share a common latent 3D radiance field. Therefore, passing these subsets of images through an ideal 3D radiance field module should yield identical latent 3D representations. Exploiting this invariance, we introduce the 3D consistency loss, formulated as:

$$L_{\text{consis}} = \frac{1}{|(X_m, X_n)| \cdot L} \sum_{(X_m, X_n)} \sum_{l=1}^{L} \|\text{Enc}_l(X_m) - \text{Enc}_l(X_n)\|_2^2 \tag{5.4}$$

where $X_m \in \mathbb{R}^{m \times H \times W \times 3}$ and $X_n \in \mathbb{R}^{n \times H \times W \times 3}$ are the two disjoint sets of images. Enc_l is the l-th encoder in the 3D radiance field module.

In an extreme scenario, the model could simply map all inputs to a single, constant representation, akin to mode collapse in generative models. However, this degenerate solution is prevented by the subsequent use of the learned representation to generate the refocused image, conditioned on the target focus map. This downstream task ensures that the 3D radiance field module produces representations that accurately capture the variations in the radiance field. To show the effectiveness of the model's capability of capturing 3D radiance, we visualize the learned 3D radiance feature in Fig 5.9.

5.2.3 Refocusing Through Focal Stack Cross-Attention Pooling

Refocus Decoder

The refocus module comprises three refocus decoders. These decoders are responsible for selectively extracting information that corresponds to the target focus map from the learned 3D representation. To accomplish this, the decoder requires a pooling mechanism capable of reducing the 3D representation to a 2D representation in a target focus-aware manner. It also needs a transformation mechanism to synthesize the refocused image from the pooled 2D features, guided by the target focus map. Figure 5.4 depicts the detailed architecture of the refocus decoder. Each decoder consists of a set of Spatial Feature Transform (SFT) blocks [215] and a novel Focal Stack Cross-Attention Pooling block.

Various methods have been proposed for conditional image-to-image translation, although their application to WSI refocusing remains unexplored. A straightforward approach involves concatenating the conditional information with intermediate feature maps, followed by a convolutional layer. However, this transformation produces a simple linear combination of the conditional information and the features. Beyond concatenation, some

methods utilize hyper-networks to generate convolutional layer weights based on the conditional information [216]. We adopt the Spatial Feature Transform (SFT) method [215], which generates affine transformation parameters for spatial feature modulation based on conditional inputs. SFT offers greater representational power compared to naive concatenation and exhibits improved stability and robustness compared to hyper-networks. In our proposed model, SFT layers are incorporated into both the encoder and decoder. Each decoder adaptively pools the 3D representation from the encoder into a 2D representation and concatenates it with the output from the previous decoder.

Focal Stack Cross Attention Pooling

Pooling operations typically reduce feature dimensions. 2D spatial pooling, commonly employed in CNNs, reduces the spatial resolution of feature maps. In many-to-one image translation tasks, such as multi-focus fusion and AIF image generation, reducing the number of images is necessary. In the context of virtual refocusing, we need to reduce the depth dimension. While max pooling and average pooling are widely used, they disregard the internal structure of the 3D radiance representation. Since the learned 3D representation is independent of the target focus map, selecting the maximum value without considering the target can only identify the most salient and representative features. While beneficial for classification tasks that do not require conditional information, this may not yield the most relevant features for refocusing the input to a specific focus map. Similarly, average pooling, which treats all information equally, remains suboptimal due to its lack of adaptability to the target focus map.

In order to make the pooling aware of the focus map condition and smartly select the features necessary for the target generation, we propose a novel adaptive pooling method called Focal Stack Cross-Attention Pooling (FSCA Pool). This method leverages a cross-attention mechanism along the depth dimension to determine the most relevant input information with respect to the target focus map f_t. The input features are then merged along the depth dimension, weighted by their relevance to the target focus map. This process is carried out in a pixel-wise manner and can be formulated as

$$
\begin{aligned}
\text{FSCA-Pool}(fea, f_t) \in \mathbb{R}^{h\times w\times c} &= \text{softmax}(\frac{\langle Q, K\rangle}{\|Q\|_2\|K\|_2}) \cdot V \\
Q \in \mathbb{R}^{1\times h\times w\times d} &= \text{net}_{\text{Q}}(f_t) \\
K \in \mathbb{R}^{D\times h\times w\times d} &= \text{net}_{\text{K}}(fea) \\
V \in \mathbb{R}^{D\times h\times w\times c} &= \text{net}_{\text{V}}(fea)
\end{aligned}
\tag{5.5}
$$

where $fea \in \mathbb{R}^{D\times h\times w\times c}$ is the 3D radiance representation and $f_t \in \mathbb{R}^{H\times W}$ is the target focus map. d is the feature dimension of the cross-attention operation. $\langle\cdot,\cdot\rangle$ is the inner product operation in the channel dimension. $\cdot$ is the elementwise multiplication. $\text{net}_{\text{Q}} : \mathbb{R}^{H\times W} \rightarrow \mathbb{R}^{1\times h\times w\times d}$ is the target focus map transformation network that consists of a convolutional layer and 2D interpolation operation. $\text{net}_{\text{K}} : \mathbb{R}^{D\times h\times w\times c} \rightarrow \mathbb{R}^{D\times h\times w\times d}$ is the query feature transformation network that consists of a single convolutional layer. $\text{net}_{\text{V}} : \mathbb{R}^{D\times h\times w\times c} \rightarrow \mathbb{R}^{D\times h\times w\times c}$ is the value feature transformation network that consists of a single convolutional layer that preserves spatial resolution of the input feature. By building the cross-attention, the query, which corresponds to the target focus map f_t, attends to the transformed input feature by comparing its similarity w.r.t. each of the inputs in the depth dimension. The similarity is represented as the gathered attention $\text{softmax}(\frac{\langle Q,K\rangle}{\|Q\|_2\|K\|_2}) \in \mathbb{R}^{D\times h\times w}$. Finally, the cross-attention is applied to the transformed input feature V to select and fuse the features that are most relevant to the target focus map f_t. Although this module only takes depth-wise attention into account, spatial-wise feature manipulation is achieved using 2D convolutions. A visualization of the attention map is shown in Fig 5.10.

5.2.4 Whole Slide Image Perceptual Distance Metric

Deep learning-based image reconstruction relies heavily on image distance metrics to guide model training. While metrics like MSE and MAE are commonly used as loss functions, they suffer from a key limitation: they treat all errors equally, assuming spatial independence. This assumption misaligns them with the characteristics of the HVS. To address this, many FR IQA models prioritize errors that are more perceptually salient to the human eye. For instance, SSIM [13] emphasizes errors that disrupt local image structure, while

neglecting non-structural errors like luminance and contrast shifts. From an optimization perspective, while these metrics share the same global optimum, those better aligned with the HVS are more likely to find superior local optima. However, a common shortcoming of these traditional FR IQA metrics is their focus on low-level visual features. They often ignore semantic information, which plays an important role in human perception of image quality and similarity.

Deep learning-based FR IQA models [11, 12] have emerged to incorporate mid-level and high-level features, along with low-level features, in assessing image quality and similarity. These models typically compute distances between features extracted at multiple levels of a pre-trained CNN. Such metrics are often referred to as perceptual distance or perceptual loss. Perceptual loss has been successfully employed in training WSI deblurring and interpolation models, yielding more realistic images compared to using MSE or MAE as loss functions. Moreover, perceptual loss tends to produce fewer artifacts compared to traditional FR IQA metrics.

Despite their advantages, a key limitation of current perceptual loss functions lies in their pre-training. For instance, both LPIPS [11] and DISTS [12] utilize a VGG16 network [210] trained on ImageNet [211]. As detailed in Section 3.2, WSIs differ significantly from natural images in several aspects, including the objects captured, the imaging devices used, illumination conditions, and post-processing techniques. While perceptual loss excels in evaluating natural image similarity, it may struggle when applied to WSIs due to this domain mismatch. To overcome this limitation, we introduce a novel perceptual distance metric specifically designed for WSIs. Our results demonstrate that incorporating this metric as the loss function substantially improves the performance of WSI refocusing compared to using natural image-based perceptual loss.

Before delving into the proposed metric, we provide a brief overview of LPIPS [11] and DISTS [12], analyzing their strengths and weaknesses to motivate the design of our WSI-specific metric. Both LPIPS and DISTS are defined as weighted sums of distances between features extracted at different layers of a pre-trained VGG16 network [210]. They can be formulated as follows:

$$D(x,y) = \sum_{l=1}^{L} \frac{1}{H_l W_l C_l} \sum_{h=1}^{H_l} \sum_{w=1}^{W_l} \sum_{c=1}^{C_l} \omega_c^l \cdot d(f(x)_{hwc}^l, f(y)_{hwc}^l) \tag{5.6}$$

where x and y are two images, f is a pretrained VGG16 network with L layers and $f(\cdot)^l \in \mathbb{R}^{H_l \times W_l \times C_l}$ is the extracted feature at layer l. The main difference between LPIPS and DISTS is the choice of the weights ω and the distance measure d. In LPIPS, $d(x,y) = \|x-y\|_2^2$, which is a simple MSE. The learnable weights are identical for all channels in the same layer, i.e., $\omega_c^l = \omega^l$. The overall formulation of LPIPS can be expressed as

$$D(x,y) = \sum_{l=1}^{L} \omega^l \|f(x)^l - f(y)^l\|_2^2 \tag{5.7}$$

Instead of measuring feature distance using MSE, DISTS uses a metric similar to SSIM, which separately measures the impact of structural and nonstructural errors. It is formulated as

$$d(x,y) = 1 - \alpha \cdot \frac{2\mu_x \mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} - \beta \cdot \frac{2\sigma_{xy} + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} \tag{5.8}$$

where μ_x, σ_x and σ_{xy} are the mean of x, standard deviation of x and covariance of x and y. α and β are the learnable weights. The overall formulation of DISTS can be expressed as

$$D(x,y) = 1 - \sum_{l=1}^{L} \sum_{c=1}^{C_l} \alpha_c^l \cdot \frac{2\mu_{xc}^l \mu_{yc}^l + c_1}{{\mu_{xc}^l}^2 + {\mu_{yc}^l}^2 + c_1} + \beta_c^l \cdot \frac{2\sigma_{xyc}^l + c_2}{{\sigma_{xc}^l}^2 + {\sigma_{yc}^l}^2 + c_2} \tag{5.9}$$

where α_c^l and β_c^l are the learnable weights of the c^{th} channel in the l^{th} convolutional layer, subject to $\sum_{l=1}^{L} \sum_{c=1}^{C_l} \alpha_c^l + \beta_c^l = 1$.

Both LPIPS and DISTS utilize features extracted from the same five layers of the VGG16 network, namely conv1_2, conv2_2, conv3_3, conv4_3, and conv5_3. By incorporating features from both shallow and deep layers, these metrics capture a wide spectrum of visual information, including low-level features like edges and textures, as well as high-level

semantic features. The key distinction lies in DISTS's inclusion of pixel-domain distances, computed without any feature extraction from the VGG16 network. This pixel-domain comparison is named as the 0^{th} layer. In contrast to LPIPS, which may not have a unique minimum, this additional pixel-domain distance ensures that DISTS has a unique minimum at $x = y$. Moreover, it has been demonstrated that $\sqrt{D(x, y)}$ in Eq. 5.9 satisfies the properties of non-negativity, symmetry, and the triangle inequality, making it a valid metric [12]. In both models, the pre-trained VGG16 network remains fixed. The only trainable parameters are the weighting scalars: $\boldsymbol{\omega} = \{\omega^l | l = 1, \cdots, 5\}$ in LPIPS and $\boldsymbol{(\alpha, \beta)} = \{(\alpha_c^l, \beta_c^l) | l = 0, \cdots, 5; c = 1, \cdots, C_l\}$ in DISTS. These parameters are learned through regression on subjectively rated image quality assessment datasets. Beyond its role as an FR IQA model, DISTS is also a texture similarity metric. Traditional FR IQA metrics often assign large feature differences to images with similar textures, despite their perceptual similarity to human observers. To address this texture invariance inherent in the HVS, DISTS incorporates an additional loss term during training to minimize Eq. 5.9 for image pairs sharing the same texture. The loss function of DISTS can be written as

$$\arg\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \frac{1}{|\mathcal{D}_1|} \sum_{x,y \in \mathcal{D}_1} \|D(x, y|\boldsymbol{\alpha}, \boldsymbol{\beta}) - q\|_1 + \gamma \cdot \frac{1}{|\mathcal{D}_2|} \sum_{z_1,z_2 \in \mathcal{D}_2} \|D(z_1, z_2|\boldsymbol{\alpha}, \boldsymbol{\beta})\|_1 \tag{5.10}$$

where (x, y) is a pair of pristine and distorted images sampled from a subjectively rated dataset $\mathcal{D}_1$. q is the MOS of y. (z_1, z_2) is a pair of texture images with the same type of texture, sampled from a texture dataset $\mathcal{D}_2$. γ is a hyperparameter that controls the tradeoff between the image quality and texture similarity.

While LPIPS and DISTS effectively assess natural image distances, their feature extractors are trained solely on natural images. Given the domain gap between WSIs and natural images in terms of both low-level statistics and high-level semantic content, employing a feature extractor specifically trained on WSI data becomes crucial. Furthermore, the texture invariance property of DISTS, while beneficial for certain natural image applications, is undesirable in medical imaging. Enforcing texture invariance in an image-to-image model can lead to significant pixel-level differences between generated images and ground truth images in textured regions. In WSI refocusing, precise pixel-level matching is important.

Hallucinated details are strictly prohibited, as they can potentially impact downstream clinical applications such as diagnosis.

To address the need for a WSI-specific image distance metric, we retrain the DISTS model on a WSI dataset, focusing on image distance assessment rather than image quality, and removing the texture invariance constraint. We term this model WSI-DISTS. Specifically, we first fine-tune the VGG16 network using the Kimia Path24 Dataset [9]. This dataset comprises 24 in-focus WSIs representing different body parts, selected from a pool of 350 slides. These slides exhibit diverse textures and visual patterns, stained using three different techniques: IHC, H&E, and Masson's trichrome staining. The WSIs were captured using a Huron TissueScope LE1.0 with a 0.75 NA lens at 20X magnification. We extracted 22,591 training patches and 1,325 validation patches of size 1000×1000 from these 24 WSIs, with representative examples shown in Figure 5.5. During fine-tuning, random rescaling was applied to enhance the model's generalization capabilities across different magnification levels. The final layer of the VGG16 network was replaced with a fully connected layer containing 24 neurons.

Having fine-tuned the VGG16 network on WSI data, we obtain a feature extractor better suited to WSI analysis than its natural image counterpart. The next step is to train the quality-related weights $\boldsymbol{\alpha}, \boldsymbol{\beta}$ using distorted WSIs. To our knowledge, there are no publicly available WSI quality assessment datasets. Therefore, we utilize the FocusPath dataset [2], which contains paired in-focus and out-of-focus WSIs with corresponding ground truth quality scores. The WSIs in FocusPath were acquired using a Huron TissueScope LE1.2 scanner [77] with a 40X objective lens at a resolution of $0.25\mu m$/pixel. A z-stack approach was employed, capturing WSIs at 16 focus levels with an increment of $0.25\mu m$. The dataset comprises 8,640 patches of size 1024×1024, extracted from 540 positions across nine different organ types and eight staining methods. Section 2.1.2 provides a more detailed description of the dataset. The training loss of WSI-DISTS is defined as

$$\arg\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \frac{1}{|\mathcal{D}|} \sum_{y_1,y_2 \in \mathcal{D}} \|D(y_1, y_2|\boldsymbol{\alpha}, \boldsymbol{\beta}) - |q_{y_1} - q_{y_2}|\|_1 \tag{5.11}$$

where (y_1, y_2) is a pair of WSI patches captured at the same location, but at different focus levels: q_{y_1} and q_{y_2}, respectively. In contrast to the original DISTS training procedure, our

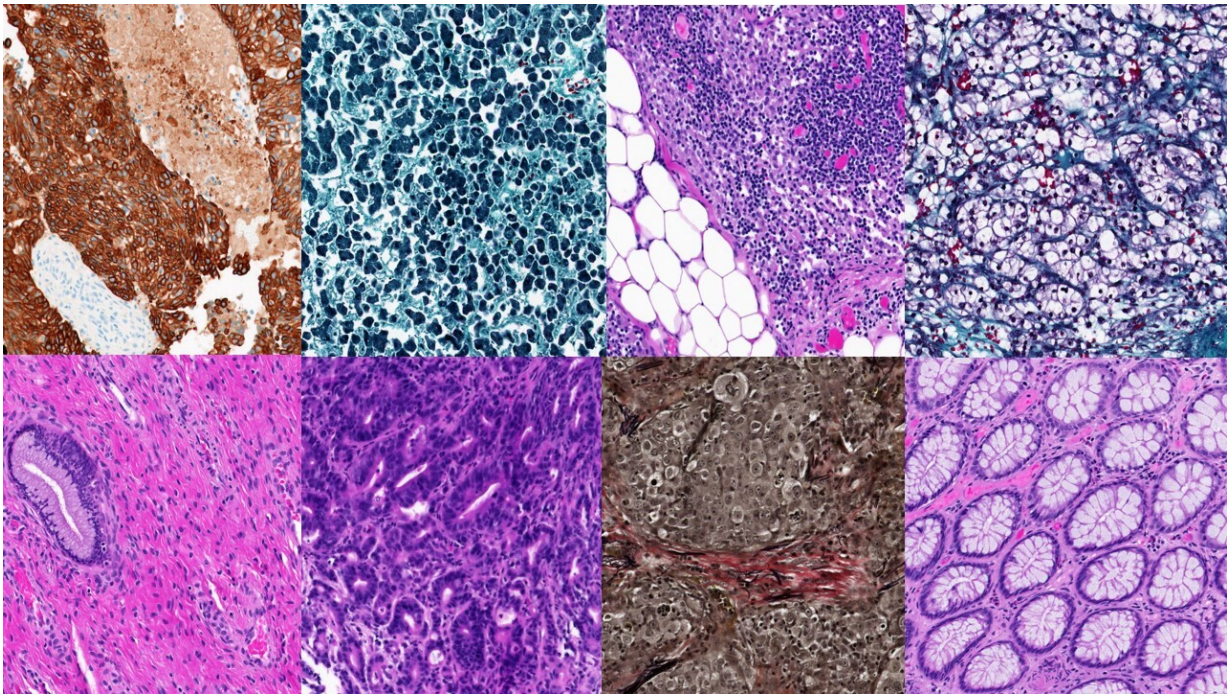


Figure 5.5: Example images of the Kimia Path24 Dataset dataset. Image credit: [9].

loss function does not require one of the inputs to $D(\cdot, \cdot)$ to be pristine. This distinction arises because DISTS is an FR IQA model, designed to assess the perceptual quality of a distorted image relative to a pristine reference. WSI-DISTS, on the other hand, aims to measure the perceptual distance between two arbitrary inputs, not necessarily involving a pristine reference. Furthermore, as texture invariance is undesirable in our application, we omit the texture invariance loss term from Eq. 5.10.

5.2.5 Overall Objective

The model is trained using supervised learning. The input images and the corresponding target image are randomly sampled from the same focal stack. The loss function consists of three components: a reconstruction loss (L_{recon}), a 3D consistency loss (L_{consis}), and an adversarial training loss (L_{GAN}). We employ the proposed WSI-DISTS metric as the reconstruction loss, defined in Eq. 5.9, with the weights obtained as described in Eq. 5.11.

The 3D consistency loss, presented in Eq. 5.4, exploits the inherent invariance of the 3D radiance field to guide the 3D radiance field module in learning a meaningful 3D tissue representation. To encourage the generation of more realistic images, we incorporate an adversarial training loss using the Least Squares GAN (LSGAN) implementation [217]. L_{GAN} is defined as

$$\begin{aligned} \hat{x_t} &= G(X, F, f_t) \\ \min_D L_{\text{GAN}_D} &= \frac{1}{2}\mathbb{E}_{(x_t, f_t)}\left[(D(x_t|f_t) - 1)^2\right] + \frac{1}{2}\mathbb{E}_{(\hat{x_t}, f_t)}\left[(D(\hat{x_t}|f_t))^2\right] \\ \min_G L_{\text{GAN}_G} &= \frac{1}{2}\mathbb{E}_{(\hat{x_t}, f_t)}\left[(D(\hat{x_t}|f_t)) - 1)^2\right] \end{aligned} \tag{5.12}$$

where G is the proposed virtual refocusing model and $\hat{x_t} = G(X, F)$ is the refocused image. $X \in \mathbb{R}^{FS \times H \times W \times 3}$ and $F \in \mathbb{R}^{FS \times H \times W \times 1}$ are the input focal stack and their corresponding focus maps, respectively. $x_t \in \mathbb{R}^{H \times W \times 3}$ and $f_t \in \mathbb{R}^{H \times W \times 1}$ are the target image and its corresponding focus map, respectively. D is a conditional discriminator.

To summarize, the overall objective for the virtual refocusing model can be written as

$$L = \lambda_r L_{\text{recon}} + \lambda_c L_{\text{consis}} + L_{\text{GAN}_G} \tag{5.13}$$

where λ_r and λ_c are the weighting parameters.

5.3 Experiments

5.3.1 Implementation Details

As discussed in Section 2.1.2, DeepFocus [1] and FocusPath [2] are the only two publicly available z-stack datasets. The image dimensions in DeepFocus (64×64) are insufficient to represent detailed tissue structures. Therefore, we utilize the FocusPath dataset, which contains images of size 1024×1024, for model training. The FocusPath dataset comprises nine slides. For each slide, two strips are scanned, and within each strip, 30 locations are

selected for focal scanning at 16 focus levels. This yields a total of 540 focal stacks, each containing 16 images of size 1024×1024, resulting in a total of 8,640 images. Figure 2.5 presents sample images from different slides, and Figure 2.4 shows an example focal stack. We employ the focal stacks from the first strip for training and those from the second strip for testing. These strips are spatially separated on the slides, ensuring minimal overlap in tissue patterns. The training dataset consists of 270 focal stacks, each with 16 images captured at different z-levels. The testing dataset also contains 270 focal stacks, each with 16 images.

The proposed virtual refocusing model accepts an arbitrary number of input images. However, for computational efficiency during training, we fix $FS = 3$. Notably, the trained model can still be evaluated with different FS values during testing. Using a larger FS generally leads to improved quality in the refocused images, as a larger number of input images contributes to a more accurate and comprehensive 3D radiance representation. When selecting the FS images from the 16 images in a focal stack, we deliberately exclude images near the target focus level. This prevents the model from learning a trivial solution that simply selects the input image closest to the target.

The model is trained end-to-end using an Adam optimizer [218] with a learning rate of 1e$-$4 and weight decay of 1e$-$7 for both the 3D radiance field module and the refocus module. The discriminator is optimized separately using another Adam optimizer with the same learning rate and weight decay. In the overall loss function (Eq. 5.13), the hyperparameters are set to $\lambda_r = 15$ and $\lambda_c = 1$. These values are not fine-tuned and serve as a reasonable starting point for optimization.

5.3.2 Evaluation Results

Figure 5.6 showcases the focus traverse capability of our virtual refocusing model. We use Target 6 and Target 11 as inputs, with the sharpest focus level occurring around level 9. For comparison with the ground truth, we employ uniform focus maps aligned with the ground truth levels in this experiment, although the target focus map can be non-uniform during both training and testing. Refocused images are generated by setting the target focus maps to uniform values ranging from 1 to 16. In the results, the refocused

images are of high quality and are perceptually similar to the ground truth images across all focus levels. The transitions between intermediate focus levels are smooth and visually consistent. Notably, this demo focal stack is from the testing dataset.

Furthermore, our experiments reveal the model's ability to refocus to target levels not encountered during training. Despite the dataset containing only 16 discrete focus levels, the model can generate smoothly refocused images using a continuous range of focus levels. This capability is illustrated in Figure 5.6, where Target 9 and Target 10 serve as inputs. The refocused images are generated by uniformly setting the target focus maps to 9.2, 9.4, 9.6, and 9.8, respectively. Upon closer inspection, the transitions between these images are remarkably smooth. It is important to highlight that this smooth transition capability is not limited to interpolation scenarios. The model can achieve similar smoothness even when using a single input image.

Next, we demonstrate the model's robustness in an extreme case where we choose the most blurry image as the single input to our model. An example is shown in Fig 5.7. The input is a single image at focus level 16, which is the most out-of-focus image in the focal stack. The refocused images are generated by setting the target focus maps uniformly from 1 to 16, respectively. It is not surprising that the qualities of generated images are inferior compared to those ones in Fig 5.6. The reason is that the input image (Target 16) lacks sufficient 3D information on the focal stack, which makes the learned 3D representation incapable of representing the real 3D radiance field. As a result, artifacts are present in refocused images at sharper levels of focus (Refocused 9).

To further evaluate model robustness, we present an extreme case where the most out-of-focus image from the focal stack serves as the only input to our model. Figure 5.7 illustrates this scenario. The input is a single image at focus level 16, representing the most blurred image in the focal stack. Refocused images are generated by setting the target focus maps uniformly to values ranging from 1 to 16. As expected, the quality of the generated images is lower compared to those shown in Figure 5.6. This degradation stems from the limited 3D information present in the blurry input image (Target 16). Consequently, the learned 3D representation cannot adequately capture the true 3D radiance field. Nevertheless, the proposed model can still produce reasonable results.

The above experiments all employed uniform focus maps, implying that the refocused image is "captured" on a single focal plane perpendicular to the optical axis. This setup simulates the imaging process in optical microscopes and WSI scanners. However, when dealing with uneven tissue surfaces, plane-wise refocusing cannot produce an image where all regions are simultaneously in focus (AIF image). Although a generated AIF image might not be desirable in all applications, we showcase our model is capable of pixel-wise refocusing. AIF generation is just a special case. This capability is demonstrated in Figure 5.8, where Target 1 and Target 8 serve as inputs. Both input images exhibit partial out-of-focus blur due to physical artifacts on the slide. The horizontal strip at the bottom is thicker than other regions. The refocused image is generated using a non-uniform focus map, shown in the third image. The result clearly shows an all-in-focus image. By enabling pixel-wise refocusing, our virtual refocusing model surpasses the limitations of traditional optical microscopy. The focus map in Figure 5.8 is manually generated. However, focus maps can also be generated automatically by adjusting the output of FQA models like FocusLiteNN [176].

To quantitatively evaluate our model, we compute four FR IQA metrics between the refocused image and the corresponding target image: SSIM [13], MS-SSIM [45], IW-SSIM [4], and DISTS [12]. Input images are randomly sampled from the focal stack. We also assess the impact of the number of input images on refocusing performance, with results presented in Table 5.3. As expected, increasing the number of input images leads to improved refocusing performance. A larger number of input images provides a richer and statistically more comprehensive representation of the 3D radiance field, enabling the 3D radiance field module to learn a better representation. This improved 3D representation, in turn, benefits the refocus module, resulting in higher-quality refocused images. Notably, even with a single input image, our model achieves reasonably good results.

Since there are no WSI refocusing models we can compare to, we evaluate our model's performance against state-of-the-art deep learning-based deblurring models: DRBNet [219], Restormer [220], and MPRNet [221]. Deblurring can be considered as a special case of the virtual refocusing model, where we only have global out-of-focus. For testing, we select the sharpest image within each focal stack as the target and use the remaining images in the stack as input, one at a time. This procedure allows us to thoroughly assess deblur-

Model	FS	SSIM ↑	MS-SSIM ↑	IW-SSIM ↑	DISTS ↓
VF@Refocus	5	0.9646	0.9603	0.9655	0.0670
VF@Refocus	4	0.9599	0.9558	0.9597	0.0709
VF@Refocus	3	0.9518	0.9482	0.9493	0.0777
VF@Refocus	2	0.9365	0.9341	0.9294	0.0895
VF@Refocus	1	0.8944	0.8980	0.8732	0.1203

Table 5.3: Refocusing performance with varying numbers of input images. Using more input images leads to better performance, as they capture richer 3D information about the radiance field.

Model	FS	SSIM ↑	MS-SSIM ↑	IW-SSIM ↑	DISTS ↓
VF@Refocus	2	0.8320	0.9273	0.9303	0.1085
VF@Interpolation	2	0.8556	0.9368	0.9462	0.1004
DFI [161]	2	0.7937	0.8237	0.7896	0.3304
LinearLatent [163]	2	0.8698	0.8775	0.8315	0.2313

Table 5.4: Focus interpolation comparison results. The proposed model outperforms the other two WSI focus interpolation models.

ring performance across a wide range of blur levels. The results are shown in Table 5.5. Remarkably, despite not being explicitly trained for deblurring, our model (VF@Refocus) outperforms the three dedicated deblurring models. We attribute this superior performance to the implicit learning of the 3D radiance field. Furthermore, utilizing three focal stack images as input further enhances performance, as this provides richer information about the 3D radiance field. We also finetuned the refocus model in the context of deblurring. The model is named as VF@Deblur. It is clear that fine-tuning is beneficial for $FS = 1$.

To highlight the effectiveness of the 3D radiance field module in capturing radiance field-related features, we present visualizations of its internal representations. Directly visualizing these features as 2D images is challenging, as they are located in different stages and are in high-dimensional space. Instead, we indirectly visualize these features by manipulating them and observing the resulting changes in the refocused images. Our hypothesis

Model	FS	SSIM ↑	MS-SSIM ↑	IW-SSIM ↑	DISTS ↓
VF@Refocus	3	0.9420	0.9312	0.9360	0.0760
VF@Refocus	1	0.8566	0.8511	0.8277	0.1192
VF@Deblur	3	0.9414	0.9312	0.9378	0.0741
VF@Deblur	1	0.8722	0.8663	0.8491	0.1029
DRBNet [219]	1	0.7060	0.7277	0.6307	0.2665
Restormer [220]	1	0.7034	0.7259	0.6149	0.2606
MPRNet [221]	1	0.6786	0.7042	0.6074	0.2466

Table 5.5: A comparison of WSI deblurring results. The proposed model outperforms the other three deep learning-based debluring models.

is that removing specific feature layers should impair the model's ability to refocus to the corresponding depths. In this experiment, we retain only the first layer of the 3D features along the depth dimension, setting all other layers to zero. This manipulation effectively preserves radiance information from the topmost tissue layer while discarding information from deeper layers. Ideally, this modified model should be unable to refocus to other depths due to this information loss. It is crucial to emphasize that we are manipulating internal features, not the input images themselves. Figure 5.9 presents the results. The left and middle refocused images are generated using only the first layer of the 3D radiance features. Despite having different target focal planes, these images appear similar. In both cases, only the topmost tissue layer (visible as a horizontal strip in the middle of the image) is in focus, consistent with the modified features containing information solely from the top tissue layer. Upon closer examination of the middle horizontal strip across the three images, the left and middle images exhibit greater nuclear detail due to being in focus at that depth. In contrast, the right image is refocused using all layers of the 3D features, demonstrating successful refocusing to the target focal plane. The topmost tissue layer in this image is out of focus.

Figure 5.10 visualizes the attention maps generated by the FSCA Pooling module for different target focus planes. In both images, each patch represents the attention map between the target and a slice in the feature. The left and right attention maps correspond

to target focus planes 1 and 7, respectively. The visualizations clearly show that FSCA Pooling selectively extracts features from the 3D radiance representation that are most relevant to the target focus. When the target is set to 1, FSCA puts more attention to shallower features. When the target is set to 7, FSCAs puts more attention to middle layer features. Importantly, the attention mechanism operates at the pixel level, rather than the layer level, enabling precise pixel-wise refocusing. This pixel-level attention is crucial in scenarios where: (1) the target focus plane is non-uniform or (2) the WSI exhibits partial or superposition out-of-focus blur.

We evaluated the inference speed of our virtual refocusing model and compared it to two WSI interpolation methods. Table 5.6 presents the results. Due to the heavy reliance on 3D convolution and 3D self-attention in the 3D Radiance Module of our virtual refocusing model, it exhibits higher computational costs compared to the two WSI interpolation methods, which do not involve 3D operations.

Model	Time (Seconds)
DFI [161]	0.05
LinearLatent [163]	0.07
Proposed	0.52

Table 5.6: Inference speed comparison of the refocus model and WSI interpolation methods. The time reported is for generating a 512×512 patch using one NVIDIA GTX 3090 GPU.

5.3.3 Ablation Study

We also conduct ablation studies to evaluate the effectiveness of the 3D radiance field module as well as the FSCA Pooling Module. The results are presented in Table 5.7 and Table 5.8. To demonstrate the effectiveness of the 3D radiance field module, we remove the 3D consistency loss L_{consis} from the overall loss function. It can be seen that the performance is inferior compared to the full model.

To further demonstrate the effectiveness of the FSCA Pooling Module, we replace it with both max pooling and mean pooling. Table 5.8 presents these ablation results. The

Model	3D Radiance	FS	SSIM ↑	MS-SSIM ↑	IW-SSIM ↑	DISTS ↓
VF@Refocus	**yes**	3	**0.9518**	**0.9482**	**0.9493**	**0.0777**
VF@Refocus	no	3	0.9034	0.9209	0.9120	0.0993

Table 5.7: Alabtion study on the effectiveness of the 3D Radiance Module. It can be seen that using this module significantly increases the quality of the refocused images.

comparison clearly shows that FSCA Pooling significantly enhances the quality of the refocused images. This improvement can be attributed to FSCA Pool's ability to understand the 3D structure of the features, unlike max pooling, which only selects the maximum value without considering structural relationships. Additionally, FSCA Pool adaptively aggregates information based on the conditional focus map, whereas mean pooling assigns equal weights to all features.

Model	Pool	SSIM ↑	MS-SSIM ↑	IW-SSIM ↑	DISTS ↓
VF@Refocus	**FSCA**	**0.9518**	**0.9482**	**0.9493**	**0.0777**
VF@Refocus	Max	0.8470	0.8583	0.8140	0.1509
VF@Refocus	Mean	0.8394	0.8509	0.8157	0.1528

Table 5.8: Alabtion study on the effectiveness of the Focal Stack Cross-Attension Pooling module. Compared to Max and Mean pooling, FSCA Pooling takes the 3D structure of the feature and the condition into account. It can be seen that using FSCA significantly increases performance.

We also provide an ablation study on the effectiveness of the WSI-DISTS metric. It can be seen from Table 5.9 that WSI-DISTS greatly enhances the quality of the refocused images compared to the original DISTS.

Model	Loss	SSIM ↑	MS-SSIM ↑	IW-SSIM ↑	DISTS ↓
VF@Refocus	WSI-DISTS	0.9518	0.9482	0.9493	0.0777
VF@Refocus	DISTS	0.9363	0.9335	0.9292	0.0893

Table 5.9: Alabtion study on the effectiveness of the WSI-DISTS reconstruction loss. Compared to the original DISTS, WSI-DISTS is fine-tuned on WSI data, which makes it more accurate in measuring the distance between WSIs.

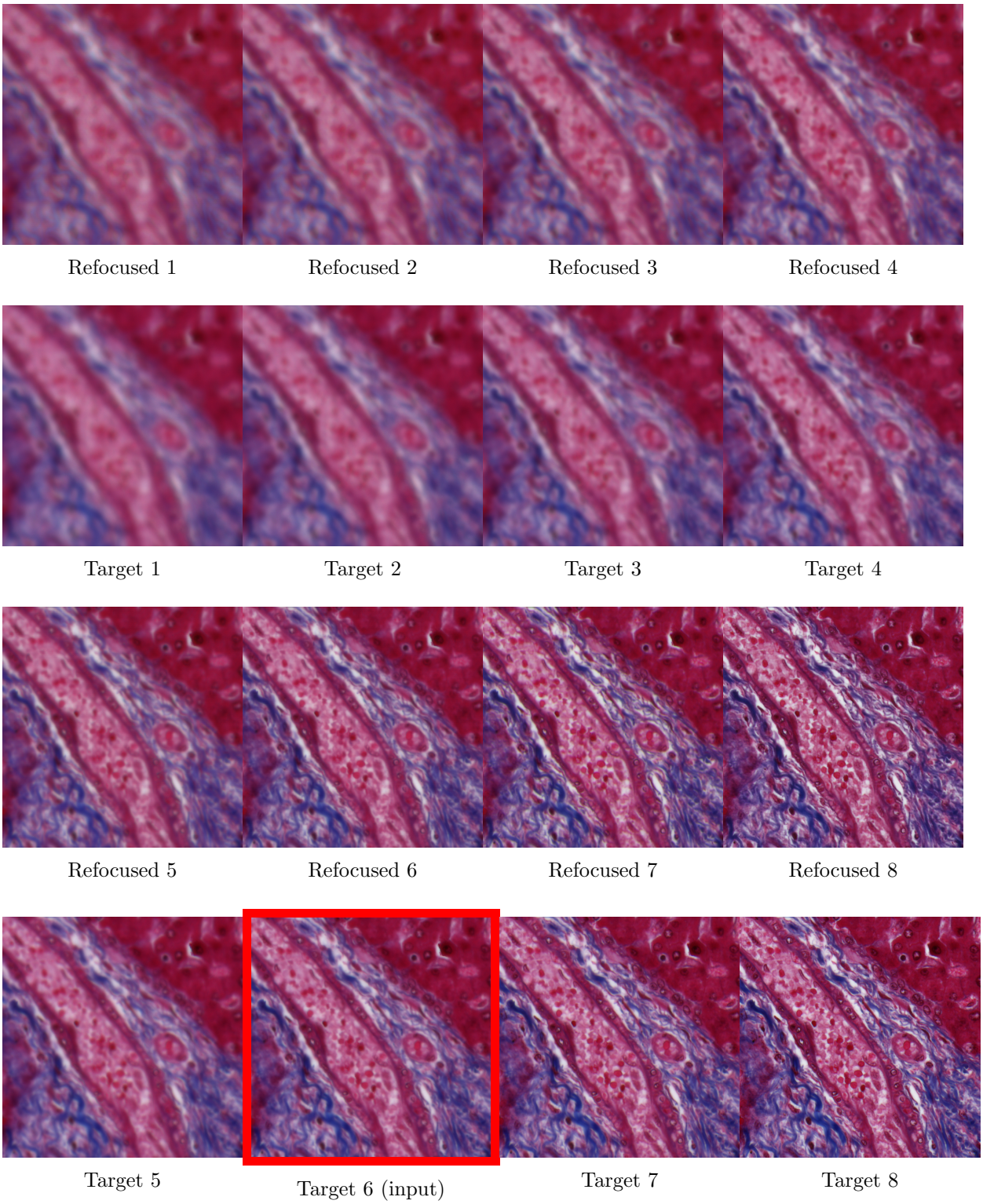
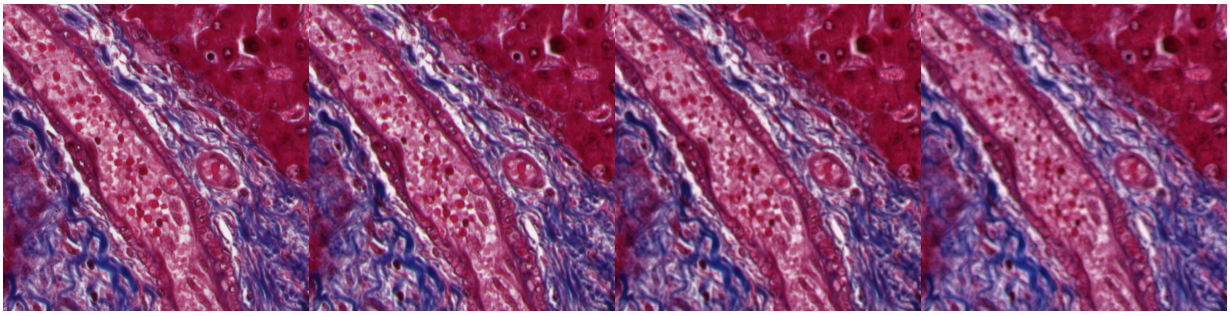
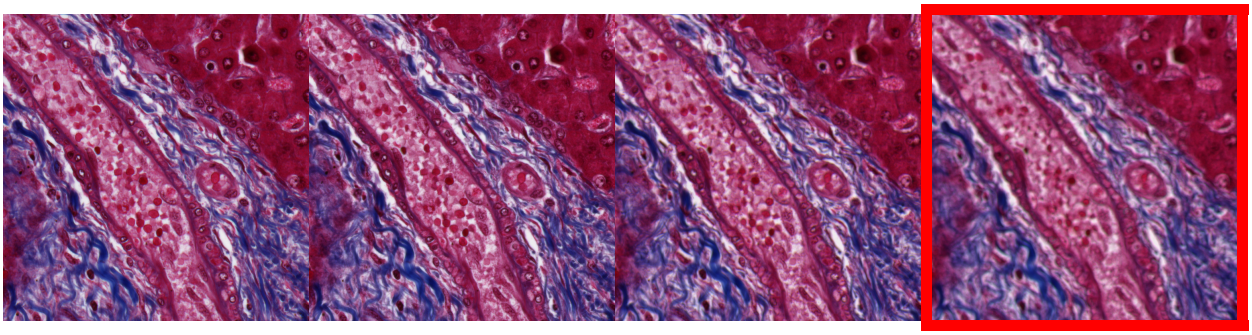


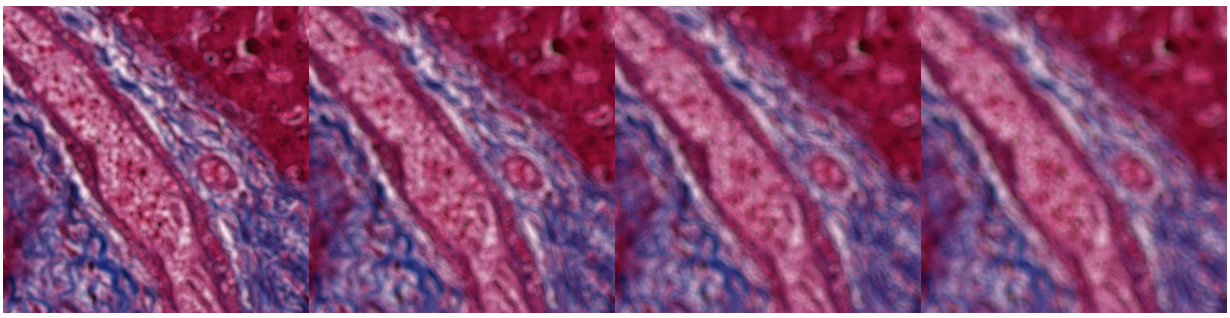
Figure 5.6: Virtual refocusing example. The inputs are Target 6 and Target 11. The refocused images are generated by setting the target focus maps to uniform ones ranging from 1 to 16.



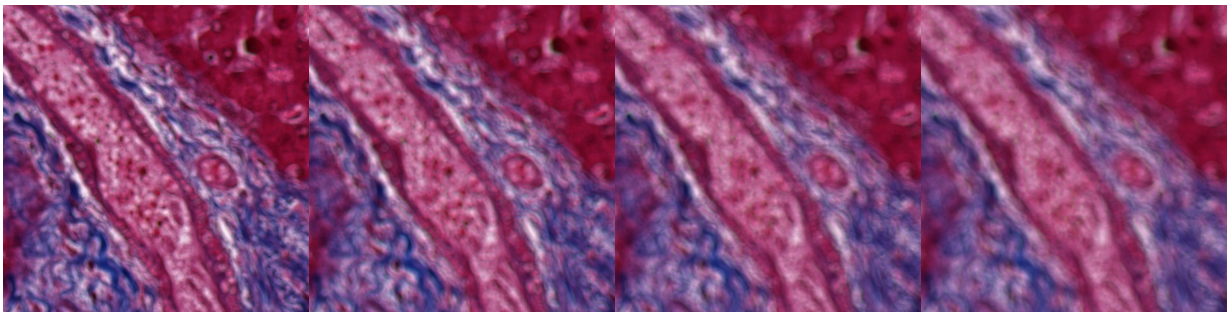
Refocused 9 | Refocused 10 | Refocused 11 | Refocused 12



Target 9 | Target 10 | Target 11 | Target 12 (input)



Refocused 13 | Refocused 14 | Refocused 15 | Refocused 16



Target 13 | Target 14 | Target 15 | Target 16

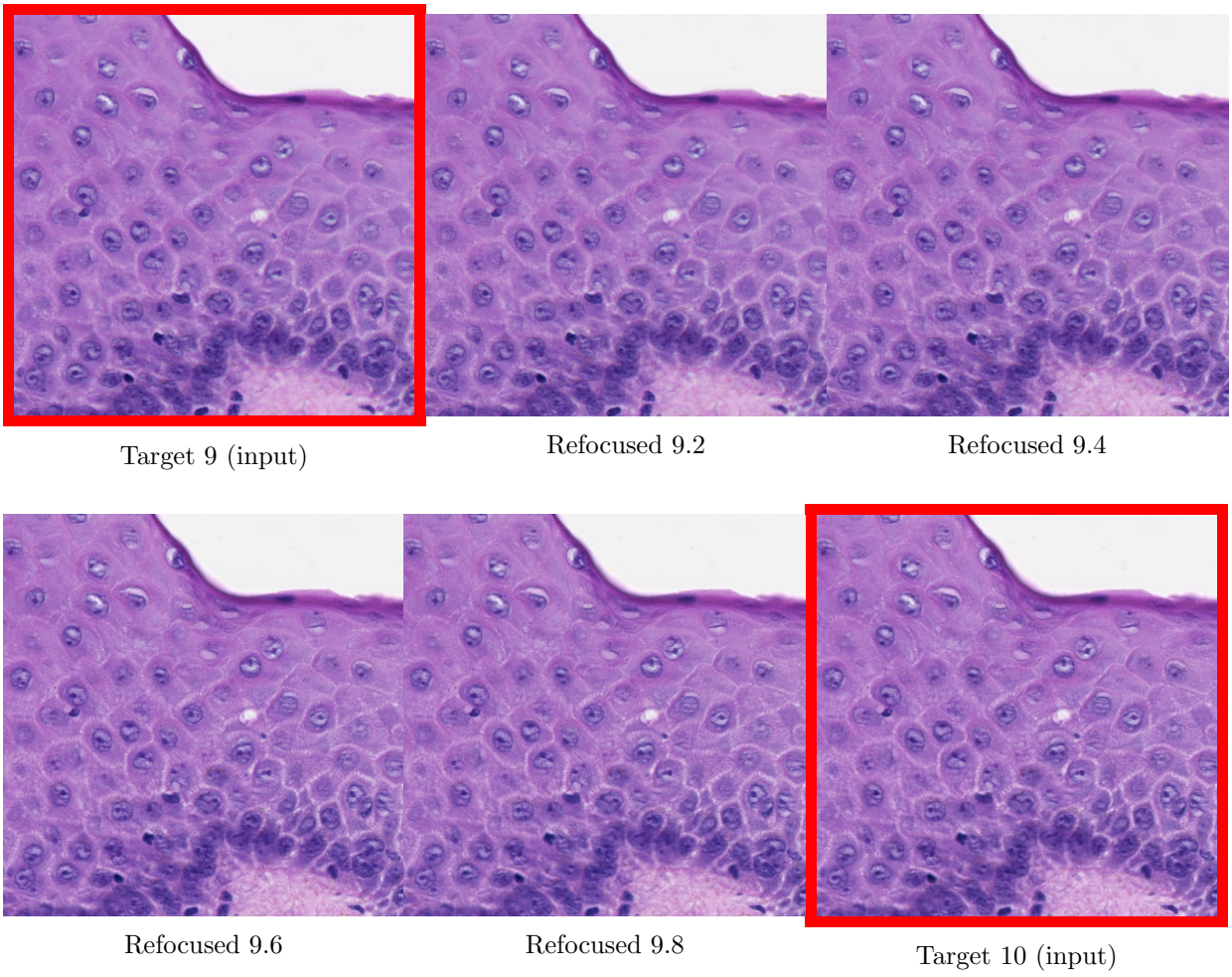
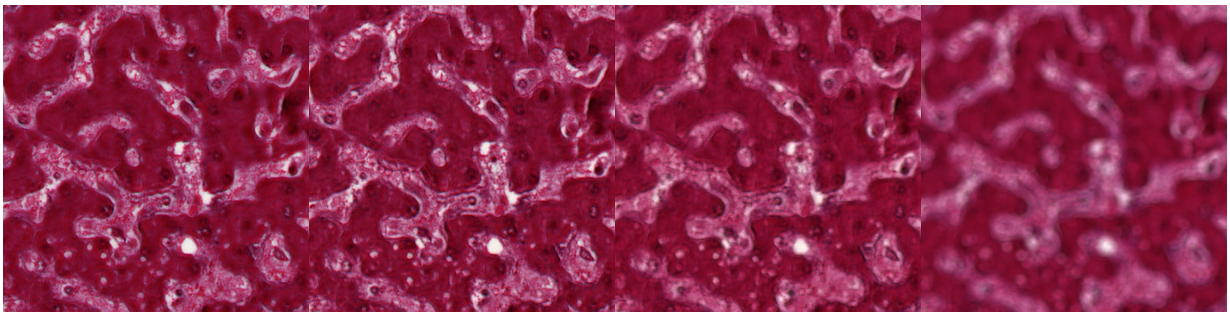
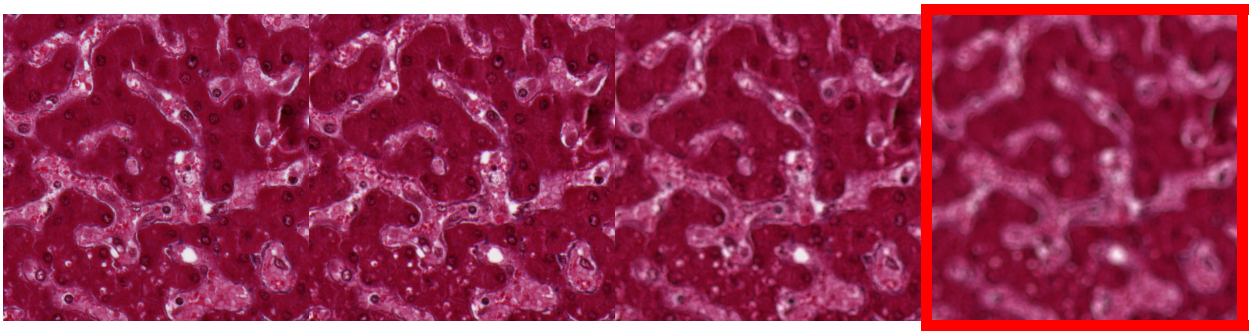


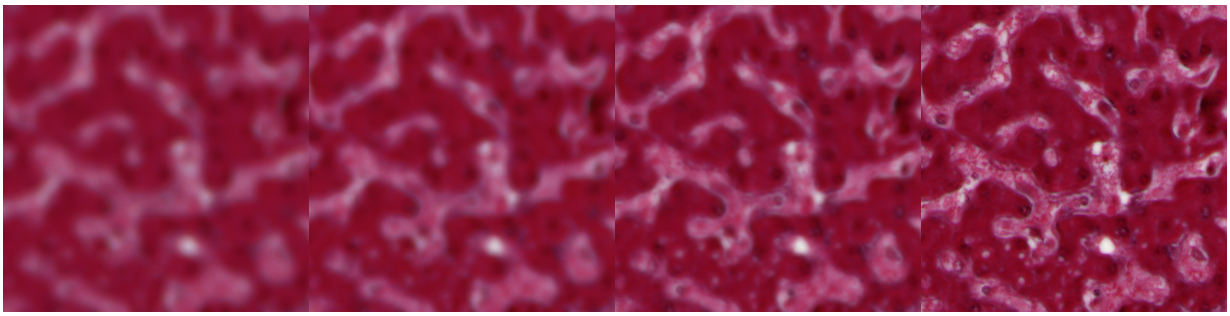
Figure 5.6: Continous refocusing example. The inputs are Target 9 and Target 10. The refocused images are generated by setting the target focus maps uniformly to 9.2, 9.4, 9.6, and 9.8, respectively.



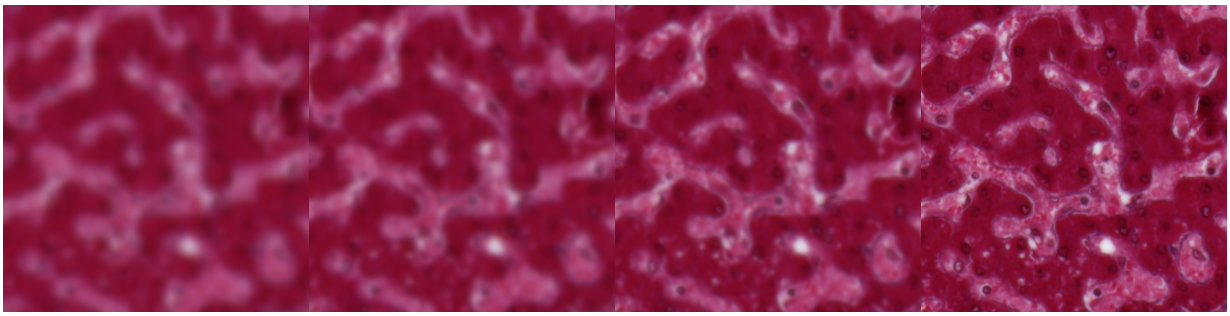
Refocused 9 Refocused 11 Refocused 13 Refocused 16



Target 9 Target 11 Target 13 Target 16 (input)

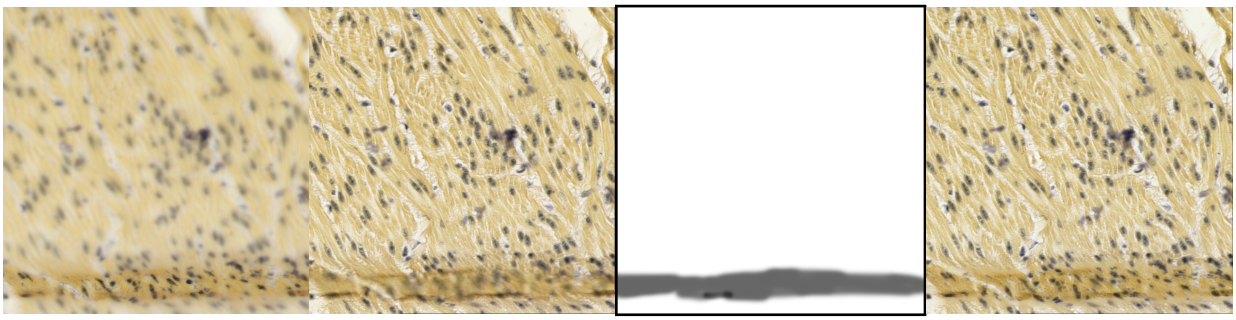


Refocused 1 Refocused 3 Refocused 5 Refocused 7



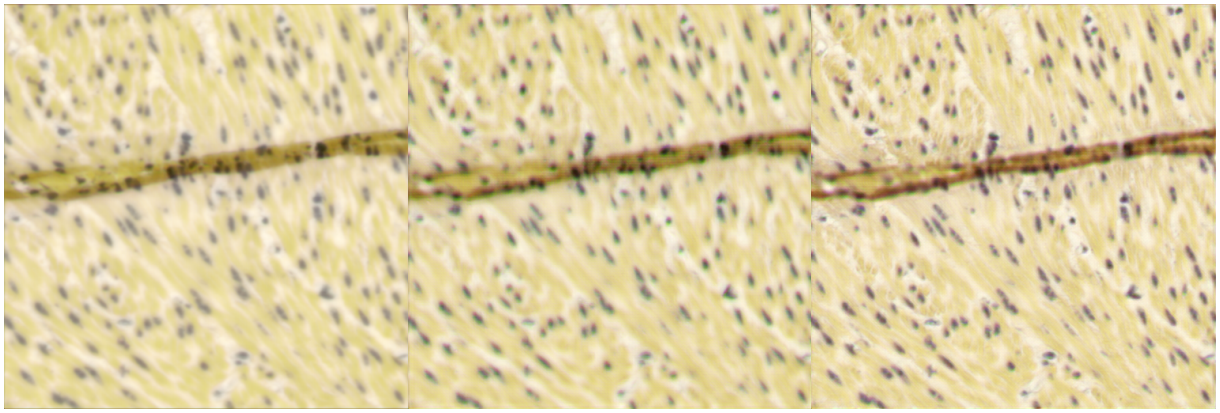
Target 1 Target 3 Target 5 Target 7

Figure 5.7: Extreme refocusing example. The input is Target 16, which is the most out-of-focus image in the focal stack. Refocused images are generated using uniform target focus maps ranging from level 1 to 16.



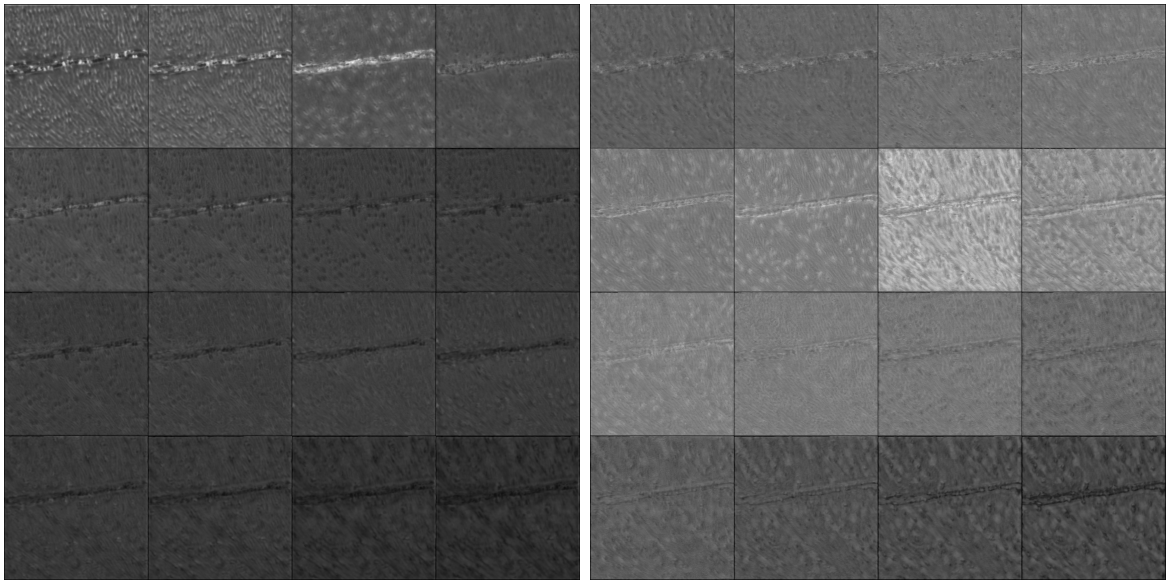
Target 1 (input) Target 8 (input) Focus Map Refocused

Figure 5.8: Non-uniform refocusing example. The inputs are Target 1 and Target 8, both exhibiting partial out-of-focus blur due to physical artifacts on the slide (note the horizontal strip at the bottom). The refocused image, generated using the non-uniform focus map shown in the third image, is in focus across all spatial locations.



Feature: 1st layer, Target: 1 Feature: 1st layer, Target: 7 Feature: all layers, Target: 7

Figure 5.9: Impact of manipulating 3D radiance features on refocusing. The left and middle images are refocused using only the first layer of the 3D radiance features, resulting in both images being focused on the topmost tissue layer despite different target focal planes. The right image, refocused using all feature layers, is correctly focused at the target focal plane.



Target: 1 Target: 7

Figure 5.10: Attention maps of the third FSCA Pooling layer. The left and right attention maps are generated when the target focus plane is set to 1 and 7, respectively. FSCA Pooling selectively attends to the layers of the 3D radiance feature representation that are most relevant to the target focus plane.

Chapter 6

Conclusion and Future Work

The first work presented in this thesis addresses the challenges of FQA for WSI. The motivation arises from the urgent need for an efficient FQA model for high-throughput WSI platforms. Based on prior knowledge about the WSIs and the imaging process, we developed the FocusLiteNN network, which contains only a single kernel. This model significantly reduces computational demands by $10,000$ times without compromising accuracy compared to SOTA network architectures. This model has been adopted in the quality control process in industry applications. Furthermore, we introduce the first open-source, expert annotated FQA dataset named TCGA@Focus. It offers a comprehensive platform for developing and evaluating new FQA models.

The second work introduced in this thesis is a IQA score fusion framework. It leverages the strengths and mitigates the weaknesses of individual IQA models by fusing their scores, resulting in a more robust model. This is achieved by incorporating both coarse-grained and fine-grained uncertainty estimation at the model level and score level, respectively. Based on MAP, this framework is the first unsupervised learning-based method for IQA score fusion. Unsupervised training allows the model to be trained on a combination of multiple datasets without the need for MOS as well as MOS alignment. Training on a large dataset improves the generalizability and reduces the model's bias.

The third project, virtual refocusing, represents a pioneering effort to address the out-of-focus problems in WSIs. This model simulates the experience of continuously adjusting

the microscope's focus, allowing for a comprehensive examination of tissue structures at varying depths without the need for physical slide presence. The input of the model is an arbitrary number of images within a focal stack. By implicitly learning a continuous 3D radiance representation from the sparse inputs, the proposed model can refocus each pixel to any focus plane according to a focus map. It also features a novel Focal Stack Cross-Attention Pooling method that gathers the information within the focal stack based on the focus map. A novel WSI distance measure WSI-DISTS is also used as the loss function to improve the performance.

The methodologies and models developed in this thesis lay a foundation for further research in several areas. Future studies of IQA for WSI can focus on more diverse distortion types other than out-of-focus blur. Developing an efficient autofocus system for WSI scanners based on FQA is also an interesting topic. The IQA score fusion work can be extended by using a more informative prior rather than uniform. The latent dimension can also be made larger than one, which gives the model more flexibility.

For virtual refocusing, a critical step is implicitly reconstructing the 3D radiance field from 2D images. This idea is similar to Neural Radiance Fields (NeRF), which can explicitly reconstruct a radiance field using images taken from various viewpoints. With adjustments to the physics involved, NeRF can also be applied to transparent objects like tissues. By incorporating NeRF into the refocusing model, the model can enhance its understanding of the 3D radiance of the tissue. Another improvement that can be made is to incorporate the knowledge of the optical system into the refocus module. The virtual refocusing model is trained on a single dataset captured with a specific optical setup. To adapt it to a new scanner, we need to retrain or fine-tune the model on a new dataset. One way to speed up the adaptation process is to incorporate PSFs of the new optical system into the refocus module. This can be achieved by initializing several convolutional layers with the PSFs and freezing them during training. When adapting to a new optical system, we can simply change the PSFs without retraining.

References

[1] C. Senaras, M. K. K. Niazi, G. Lozanski, and M. N. Gurcan, "Deepfocus: detection of out-of-focus regions in whole slide digital images using deep learning," *PloS one*, vol. 13, no. 10, 2018.

[2] M. S. Hosseini, Y. Zhang, and K. N. Plataniotis, "Encoding visual sensitivity by maxpol convolution filters for image sharpness assessment," *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4510–4525, 2019.

[3] M. Häggström, "Folding artifact on whole slide imaging of bone, causing defocus aberration (blur)." https://en.wikipedia.org/wiki/File:Folding_artifact_on_whole_slide_imaging_of_bone.png, 2023.

[4] Z. Wang and Q. Li, "Information Content Weighting for Perceptual Image Quality Assessment," *IEEE Trans. Image Process.*, 2011.

[5] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A Feature Similarity Index for Image Quality Assessment," *IEEE Trans. Image Process.*, vol. 20, pp. 2378–2386, Aug. 2011.

[6] L. Zhang, Y. Shen, and H. Li, "VSI: A Visual Saliency-Induced Index for Perceptual Image Quality Assessment," *IEEE Trans. Image Process.*, vol. 23, pp. 4270–4281, Oct. 2014.

[7] H. Lin, V. Hosu, and D. Saupe, "KADID-10k: A Large-scale Artificially Distorted IQA Database," in *Proc. Int. Conf. Qual. Multimedia Exper.)*, pp. 1–3, June 2019.

[8] A. Zarić, N. Tatalović, N. Brajković, H. Hlevnjak, M. Lončarić, E. Dumić, and S. Grgić, "VCL@FER Image Quality Assessment Database," *AUTOMATIKA*, vol. 53, no. 4, pp. 344–354, 2012.

[9] S. Shafiei, M. Babaie, S. Kalra, and H. R. Tizhoosh, "Colored kimia path24 dataset: Configurations and benchmarks with deep embeddings," 2021.

[10] "The Cancer Genome Atlas." https://en.wikipedia.org/wiki/The_Cancer_Genome_Atlas.

[11] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

[12] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2567–2581, 2022.

[13] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Trans. Image Process.*, vol. 13, pp. 600–612, Apr. 2004.

[14] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, pp. 273–297, 1995.

[15] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *European conference on computational learning theory*, pp. 23–37, Springer, 1995.

[16] F. C. Groen, I. T. Young, and G. Ligthart, "A comparison of different focus functions for use in autofocus algorithms," *Cytometry: The Journal of the International Society for Analytical Cytology*, vol. 6, no. 2, pp. 81–91, 1985.

[17] T. Yeo, S. Ong, R. Sinniah, *et al.*, "Autofocusing for tissue microscopy," *Image and vision computing*, vol. 11, no. 10, pp. 629–639, 1993.

[18] A. Santos, C. ORTIZ DE SOLÓRZANO, J. J. Vaquero, J. M. Pena, N. Malpica, and F. del Pozo, "Evaluation of autofocus functions in molecular cytogenetic analysis," *Journal of microscopy*, vol. 188, no. 3, pp. 264–272, 1997.

[19] R. A. Jarvis, "Focus optimisation criteria for computer image processing.," *Microscope*, vol. 24, no. 2, pp. 163–180, 1976.

[20] J. F. Brenner, B. S. Dew, J. B. Horton, T. King, P. W. Neurath, and W. D. Selles, "An automated microscope for cytologic research a preliminary evaluation.," *Journal of Histochemistry & Cytochemistry*, vol. 24, no. 1, pp. 100–111, 1976.

[21] D. Vollath, "Automatic focusing by correlative methods," *Journal of Microscopy*, vol. 147, no. 3, pp. 279–288, 1987.

[22] D. Vollath, "The influence of the scene parameters and of noise on the behaviour of automatic focusing algorithms," *Journal of microscopy*, vol. 151, no. 2, pp. 133–146, 1988.

[23] A. Santos, C. ORTIZ DE SOLÓRZANO, J. J. Vaquero, J. M. Pena, N. Malpica, and F. del Pozo, "Evaluation of autofocus functions in molecular cytogenetic analysis," *Journal of microscopy*, vol. 188, no. 3, pp. 264–272, 1997.

[24] L. Firestone, K. Cook, K. Culp, N. Talsania, and K. Preston Jr, "Comparison of autofocus methods for automated microscopy," *Cytometry: The Journal of the International Society for Analytical Cytology*, vol. 12, no. 3, pp. 195–206, 1991.

[25] J. F. Schlag, A. C. Sanderson, C. P. Neuman, and F. C. Wimberly, *Implementation of automatic focusing algorithms for a computer vision system with camera control*. Carnegie-Mellon University, The Robotics Institute, 1983.

[26] M. Kristan, J. Perš, M. Perše, and S. Kovačič, "A bayes-spectral-entropy-based measure of camera focus using a discrete cosine transform," *Pattern Recognition Letters*, vol. 27, no. 13, pp. 1431–1439, 2006.

[27] P. A. Devijver, "On a new class of bounds on bayes risk in multihypothesis pattern recognition," *IEEE Transactions on Computers*, vol. 100, no. 1, pp. 70–80, 1974.

[28] G. Yang and B. J. Nelson, "Wavelet-based autofocusing and unsupervised segmentation of microscopic images," in *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453)*, vol. 3, pp. 2143–2148, IEEE, 2003.

[29] G. Yang and B. J. Nelson, "Micromanipulation contact transition control by selective focusing and microforce control," in *2003 IEEE International Conference on Robotics and Automation (Cat. No. 03CH37422)*, vol. 3, pp. 3200–3206, IEEE, 2003.

[30] D. Padfield, J. Rittscher, and B. Roysam, "Defocus and low cnr detection for cell tracking applications," in *MIAAB Workshop*, Citeseer, 2008.

[31] J. M. Tenenbaum, *Accommodation in computer vision*. Stanford University, 1971.

[32] S. K. Nayar and Y. Nakagawa, "Shape from focus," *IEEE Transactions on Pattern analysis and machine intelligence*, vol. 16, no. 8, pp. 824–831, 1994.

[33] M. Subbarao, T.-S. Choi, and A. Nikzad, "Focusing techniques," *Optical Engineering*, vol. 32, no. 11, pp. 2824–2836, 1993.

[34] X. Moles Lopez, E. D'Andrea, P. Barbot, A.-S. Bridoux, S. Rorive, I. Salmon, O. Debeir, and C. Decaestecker, "An automated blur detection method for histological whole slide imaging," *PloS one*, vol. 8, no. 12, p. e82710, 2013.

[35] B. Lahrmann, N. A. Valous, U. Eisenmann, N. Wentzensen, and N. Grabe, "Semantic focusing allows fully automated single-layer slide scanning of cervical cytology slides," *PloS one*, vol. 8, no. 4, p. e61441, 2013.

[36] N. D. Narvekar and L. J. Karam, "A no-reference image blur metric based on the cumulative probability of blur detection (cpbd)," *IEEE Transactions on Image Processing*, vol. 20, no. 9, pp. 2678–2683, 2011.

[37] R. Ferzli and L. J. Karam, "A no-reference objective image sharpness metric based on the notion of just noticeable blur (jnb)," *IEEE transactions on image processing*, vol. 18, no. 4, pp. 717–728, 2009.

[38] N. Hashimoto, P. A. Bautista, M. Yamaguchi, N. Ohyama, and Y. Yagi, "Reference-less image quality evaluation for whole slide imaging," *Journal of pathology informatics*, vol. 3, no. 1, p. 9, 2012.

[39] S. Walkowski and J. Szymas, "Quality evaluation of virtual slides using methods based on comparing common image areas," in *Diagnostic pathology*, vol. 6, pp. 1–7, BioMed Central, 2011.

[40] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Transactions on systems, man, and cybernetics*, no. 6, pp. 610–621, 1973.

[41] D. Ameisen, C. Deroulers, V. Perrier, J.-B. Yunès, F. Bouhidel, M. Battistella, L. Legrès, A. Janin, and P. Bertheau, "Stack or trash? quality assessment of virtual slides," in *Diagnostic Pathology*, vol. 8, pp. 1–5, Springer, 2013.

[42] D. Ameisen, C. Deroulers, V. Perrier, F. Bouhidel, M. Battistella, L. Legrès, A. Janin, P. Bertheau, and J.-B. Yunès, "Towards better digital pathology workflows: programming libraries for high-speed sharpness assessment of whole slide images," in *Diagnostic pathology*, vol. 9, pp. 1–7, BioMed Central, 2014.

[43] A. Jiménez, G. Bueno, G. Cristóbal, O. Déniz, D. Toomey, and C. Conway, "Image quality metrics applied to digital pathology," in *Optics, Photonics and Digital Technologies for Imaging Applications IV*, vol. 9896, pp. 170–187, SPIE, 2016.

[44] Z. Wang and A. C. Bovik, "A Universal Image Quality Index," *IEEE Signal Process. Lett.*, vol. 9, pp. 81–84, Mar. 2002.

[45] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. Asilomar Conf. Signals, Syst., Comput. (ASILOMAR)*, pp. 1398–1402, Nov. 2003.

[46] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An Information Fidelity Criterion for Image Quality Assessment Using Natural Scene Statistics," *IEEE Trans. Image Process.*, vol. 14, pp. 2117–2128, Dec. 2005.

[47] H. R. Sheikh and A. C. Bovik, "Image Information and Visual Quality," *IEEE Trans. Image Process.*, vol. 15, pp. 430–444, Feb. 2006.

[48] A. K. Moorthy and A. C. Bovik, "A Two-Step Framework for Constructing Blind Image Quality Indices," *IEEE Signal Process. Lett.*, vol. 17, pp. 513–516, May 2010.

[49] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-Reference Image Quality Assessment in the Spatial Domain," *IEEE Trans. Image Process.*, vol. 21, pp. 4695–4708, Dec. 2012.

[50] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "Completely Blind" Image Quality Analyzer," *IEEE Signal Process. Lett.*, vol. 20, pp. 209–212, Mar. 2013.

[51] D. Gao, D. Padfield, J. Rittscher, and R. McKay, "Automated training data generation for microscopy focus classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 446–453, Springer, 2010.

[52] G. Campanella, A. R. Rajanna, L. Corsale, P. J. Schüffler, Y. Yagi, and T. J. Fuchs, "Towards machine learned quality control: A benchmark for sharpness quantification in digital pathology," *Computerized Medical Imaging and Graphics*, vol. 65, pp. 142–151, 2018.

[53] X. Marichal, W.-Y. Ma, and H. Zhang, "Blur determination in the compressed domain using dct information," in *Proceedings 1999 International Conference on Image Processing (Cat. 99CH36348)*, vol. 2, pp. 386–390, IEEE, 1999.

[54] H. Tong, M. Li, H. Zhang, and C. Zhang, "Blur detection for digital images using wavelet transform," in *2004 IEEE international conference on multimedia and expo (ICME)(IEEE Cat. No. 04TH8763)*, vol. 1, pp. 17–20, IEEE, 2004.

[55] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "Perceptual blur and ringing metrics: application to jpeg2000," *Signal processing: Image communication*, vol. 19, no. 2, pp. 163–172, 2004.

[56] K. De and V. Masilamani, "Image sharpness measure for blurred images in frequency domain," *Procedia Engineering*, vol. 64, pp. 149–158, 2013.

[57] 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*, vol. 39, no. 1, pp. 62–74, 2019.

[58] M. S. Hosseini and K. N. Plataniotis, "Image sharpness metric based on maxpol convolution kernels," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 296–300, Oct 2018.

[59] S. J. Yang, M. Berndl, D. M. Ando, M. Barch, A. Narayanaswamy, E. Christiansen, S. Hoyer, C. Roat, J. Hung, C. T. Rueden, *et al.*, "Assessing microscope image focus quality with deep learning," *BMC bioinformatics*, vol. 19, no. 1, p. 77, 2018.

[60] C. Senaras, M. K. K. Niazi, G. Lozanski, and M. N. Gurcan, "Deepfocus: detection of out-of-focus regions in whole slide digital images using deep learning," *PloS one*, vol. 13, no. 10, p. e0205387, 2018.

[61] T. Kohlberger, Y. Liu, M. Moran, P.-H. C. Chen, T. Brown, J. D. Hipp, C. H. Mermel, and M. C. Stumpe, "Whole-slide image focus quality: Automatic assessment and impact on ai cancer detection," *Journal of pathology informatics*, vol. 10, no. 1, p. 39, 2019.

[62] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[63] N. Wadhwa, R. Garg, D. E. Jacobs, B. E. Feldman, N. Kanazawa, R. Carroll, Y. Movshovitz-Attias, J. T. Barron, Y. Pritch, and M. Levoy, "Synthetic depth-of-field with a single-camera mobile phone," *ACM Transactions on Graphics (ToG)*, vol. 37, no. 4, pp. 1–13, 2018.

[64] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, June 2016.

[65] A. H. Murphy, "A note on the ranked probability score," *Journal of Applied Meteorology and Climatology*, vol. 10, no. 1, pp. 155–156, 1971.

[66] T. Albuquerque, A. Moreira, and J. S. Cardoso, "Deep ordinal focus assessment for whole slide images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 657–663, October 2021.

[67] E. Frank and M. Hall, "A simple approach to ordinal classification," in *Machine Learning: ECML 2001: 12th European Conference on Machine Learning Freiburg, Germany, September 5–7, 2001 Proceedings 12*, pp. 145–156, Springer, 2001.

[68] C. Beckham and C. Pal, "Unimodal probability distributions for deep ordinal classification," in *Proceedings of the 34th International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.), vol. 70 of *Proceedings of Machine Learning Research*, pp. 411–419, PMLR, 06–11 Aug 2017.

[69] T. Albuquerque, R. Cruz, and J. S. Cardoso, "Ordinal losses for classification of cervical cancer risk," *PeerJ Computer Science*, vol. 7, p. e457, 2021.

[70] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

[71] J. Liao, X. Chen, G. Ding, P. Dong, H. Ye, H. Wang, Y. Zhang, and J. Yao, "Deep learning-based single-shot autofocus method for digital microscopy," *Biomedical Optics Express*, vol. 13, no. 1, pp. 314–327, 2022.

[72] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1314–1324, 2019.

[73] M. Jaderberg, V. Dalibard, S. Osindero, W. M. Czarnecki, J. Donahue, A. Razavi, O. Vinyals, T. Green, I. Dunning, K. Simonyan, *et al.*, "Population based training of neural networks," *arXiv preprint arXiv:1711.09846*, 2017.

[74] A. Patel, U. G. Balis, J. Cheng, Z. Li, G. Lujan, D. S. McClintock, L. Pantanowitz, and A. Parwani, "Contemporary whole slide imaging devices and their applications within the modern pathology department: A selected hardware review," *Journal of Pathology Informatics*, vol. 12, no. 1, p. 50, 2021.

[75] D. A. Gutman, J. Cobb, D. Somanna, Y. Park, F. Wang, T. Kurc, J. H. Saltz, D. J. Brat, L. A. Cooper, and J. Kong, "Cancer digital slide archive: an informatics resource to support integrated in silico analysis of tcga pathology data," *Journal of the American Medical Informatics Association*, vol. 20, no. 6, pp. 1091–1098, 2013.

[76] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, July 2017.

[77] A. E. Dixon, "Pathology slide scanner," Nov. 25 2014. US Patent 8,896,918.

[78] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "Perceptual blur and ringing metrics: application to jpeg2000," *Signal processing: Image communication*, vol. 19, no. 2, pp. 163–172, 2004.

[79] C. Feichtenhofer, H. Fassold, and P. Schallauer, "A perceptual image sharpness metric based on local edge gradient analysis," *IEEE Signal Processing Letters*, vol. 20, no. 4, pp. 379–382, 2013.

[80] R. Zhang, Q. Xiao, Y. Du, and X. Zuo, "Dspi filtering evaluation method based on sobel operator and image entropy," *IEEE Photonics Journal*, vol. 13, no. 6, pp. 1–10, 2021.

[81] E. Ong, W. Lin, Z. Lu, X. Yang, S. Yao, F. Pan, L. Jiang, and F. Moschetti, "A no-reference quality metric for measuring image blur," in *Seventh International Symposium on Signal Processing and Its Applications, 2003. Proceedings.*, vol. 1, pp. 469–472, Ieee, 2003.

[82] Y. Zhan and R. Zhang, "No-reference image sharpness assessment based on maximum gradient and variability of gradients," *IEEE Transactions on Multimedia*, vol. 20, no. 7, pp. 1796–1808, 2018.

[83] J. Andrade, "No-reference image sharpness assessment based on perceptually-weighted image gradients," in *2023 14th International Conference on Information, Intelligence, Systems & Applications (IISA)*, pp. 1–8, IEEE, 2023.

[84] V. Kayargadde and J.-B. Martens, "Estimation of edge parameters and image blur using polynomial transforms," *CVGIP: Graphical models and image processing*, vol. 56, no. 6, pp. 442–461, 1994.

[85] V. Kayargadde and J.-B. Martens, "Perceptual characterization of images degraded by blur and noise: model," *JOSA A*, vol. 13, no. 6, pp. 1178–1188, 1996.

[86] J. H. Elder and S. W. Zucker, "Local scale control for edge detection and blur estimation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 7, pp. 699–716, 1998.

[87] J. Guan, W. Zhang, J. Gu, and H. Ren, "No-reference blur assessment based on edge modeling," *Journal of Visual Communication and Image Representation*, vol. 29, pp. 1–7, 2015.

[88] S. Wu, W. Lin, Z. Lu, E. P. Ong, and S. Yao, "Blind blur assessment for vision-based applications," in *2007 IEEE International Conference on Multimedia and Expo*, pp. 1639–1642, 2007.

[89] Z. Liu, H. Hong, Z. Gan, J. Wang, and Y. Chen, "An improved method for evaluating image sharpness based on edge information," *Applied Sciences*, vol. 12, no. 13, p. 6712, 2022.

[90] B. Pei, X. Liu, and Z. Feng, "A no-reference image sharpness metric based on large-scale structure," in *Journal of Physics: Conference Series*, vol. 960, p. 012018, IOP Publishing, 2018.

[91] L. Liu, J. Gong, H. Huang, and Q. Sang, "Blind image blur metric based on orientation-aware local patterns," *Signal Processing: Image Communication*, vol. 80, p. 115654, 2020.

[92] J. Chen, S. Li, L. Lin, J. Wan, and Z. Li, "No-reference blurred image quality assessment method based on structure of structure features," *Signal Processing: Image Communication*, vol. 118, p. 117008, 2023.

[93] K. Bahrami and A. C. Kot, "A fast approach for no-reference image sharpness assessment based on maximum local variation," *IEEE Signal Processing Letters*, vol. 21, no. 6, pp. 751–755, 2014.

[94] C. T. Vu, T. D. Phan, and D. M. Chandler, "s_3: a spectral and spatial measure of local perceived sharpness in natural images," *IEEE transactions on image processing*, vol. 21, no. 3, pp. 934–945, 2011.

[95] K. Gu, G. Zhai, W. Lin, X. Yang, and W. Zhang, "No-reference image sharpness assessment in autoregressive parameter space," *IEEE Transactions on Image Processing*, vol. 24, no. 10, pp. 3218–3231, 2015.

[96] K. Friston, "The free-energy principle: a unified brain theory?," *Nature reviews neuroscience*, vol. 11, no. 2, pp. 127–138, 2010.

[97] L. Li, W. Lin, X. Wang, G. Yang, K. Bahrami, and A. C. Kot, "No-reference image blur assessment based on discrete orthogonal moments," *IEEE Transactions on Cybernetics*, vol. 46, no. 1, pp. 39–50, 2016.

[98] C. Deng, S. Wang, Z. Li, G.-B. Huang, and W. Lin, "Content-insensitive blind image blurriness assessment using weibull statistics and sparse extreme learning machine," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 3, pp. 516–527, 2019.

[99] W. Xue and X. Mou, "Reduced reference image quality assessment based on weibull statistics," in *2010 Second International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, IEEE, 2010.

[100] S. Lyu and E. P. Simoncelli, "Nonlinear image representation using divisive normalization," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[101] S. Balasundaram, D. Gupta, *et al.*, "1-norm extreme learning machine for regression and multiclass classification using newton method," *Neurocomputing*, vol. 128, pp. 4–14, 2014.

[102] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 2, pp. 513–529, 2012.

[103] R. Ferzli and L. J. Karam, "No-reference objective wavelet based noise immune image sharpness metric," in *IEEE International Conference on Image Processing 2005*, vol. 1, pp. I–405, IEEE, 2005.

[104] P. V. Vu and D. M. Chandler, "A fast wavelet-based algorithm for global and local image sharpness estimation," *IEEE Signal Processing Letters*, vol. 19, no. 7, pp. 423–426, 2012.

[105] Y. Liu, G. Zhai, X. Liu, and D. Zhao, "Quality assessment for out-of-focus blurred images," in *2015 Visual Communications and Image Processing (VCIP)*, pp. 1–4, IEEE, 2015.

[106] Z. Wang and E. Simoncelli, "Local phase coherence and the perception of blur," *Advances in neural information processing systems*, vol. 16, 2003.

[107] R. Hassen, Z. Wang, and M. M. A. Salama, "Image sharpness assessment based on local phase coherence," *IEEE Transactions on Image Processing*, vol. 22, pp. 2798–2810, July 2013.

[108] A. Leclaire and L. Moisan, "No-reference image quality assessment and blind deblurring with sharpness metrics exploiting fourier phase information," *Journal of Mathematical Imaging and Vision*, vol. 52, no. 1, pp. 145–172, 2015.

[109] G. Blanchet, L. Moisan, and B. Rouge, "Measuring the global phase coherence of an image," in *2008 15th IEEE International Conference on Image Processing*, pp. 1176–1179, 2008.

[110] J. Caviedes and F. Oberti, "A new sharpness metric based on local kurtosis, edge and energy information," *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 147–161, 2004.

[111] L. Li, W. Xia, W. Lin, Y. Fang, and S. Wang, "No-reference and robust image sharpness evaluation based on multiscale spatial and spectral features," *IEEE Transactions on Multimedia*, vol. 19, no. 5, pp. 1030–1040, 2016.

[112] Q. Sang, H. Qi, X. Wu, C. Li, and A. C. Bovik, "No-reference image blur index based on singular value curve," *Journal of Visual Communication and Image Representation*, vol. 25, no. 7, pp. 1625–1630, 2014.

[113] S. Zhang, P. Li, X. Xu, L. Li, and C.-C. Chang, "No-reference image blur assessment based on response function of singular values," *Symmetry*, vol. 10, no. 8, p. 304, 2018.

[114] A. Ciancio, A. L. N. Targino da Costa, E. A. B. da Silva, A. Said, R. Samadani, and P. Obrador, "No-Reference Blur Assessment of Digital Pictures Based on Multifeature Classifiers," *IEEE Trans. Image Process.*, vol. 20, pp. 64–75, Jan. 2011.

[115] T. Oh, J. Park, K. Seshadrinathan, S. Lee, and A. C. Bovik, "No-reference sharpness assessment of camera-shaken images by analysis of spectral structure," *IEEE Transactions on Image Processing*, vol. 23, no. 12, pp. 5428–5439, 2014.

[116] L. Li, D. Wu, J. Wu, H. Li, W. Lin, and A. C. Kot, "Image sharpness assessment by sparse representation," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 1085–1097, 2016.

[117] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?," *Vision research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[118] Q. Lu, W. Zhou, and H. Li, "A no-reference image sharpness metric based on structural information using sparse representation," *Information Sciences*, vol. 369, pp. 334–346, 2016.

[119] Y. Zhang, H. Wang, F. Tan, W. Chen, and Z. Wu, "No-reference image sharpness assessment based on rank learning," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 2359–2363, IEEE, 2019.

[120] D. Li, T. Jiang, W. Lin, and M. Jiang, "Which has better visual quality: The clear blue sky or a blurry animal?," *IEEE Transactions on Multimedia*, vol. 21, no. 5, pp. 1221–1234, 2019.

[121] D. Li, T. Jiang, and M. Jiang, "Exploiting high-level semantics for no-reference image quality assessment of realistic blur images," in *Proceedings of the 25th ACM international conference on Multimedia*, pp. 378–386, 2017.

[122] L. Li, Y. Zhou, K. Gu, Y. Yang, and Y. Fang, "Blind realistic blur assessment based on discrepancy learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 11, pp. 3859–3869, 2020.

[123] C. Guo, Z. Bian, S. Alhudaithy, S. Jiang, Y. Tomizawa, P. Song, T. Wang, and X. Shao, "Brightfield, fluorescence, and phase-contrast whole slide imaging via dual-led autofocusing," *Biomedical Optics Express*, vol. 12, no. 8, pp. 4651–4660, 2021.

[124] 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*, vol. 13, no. 12, p. e202000227, 2020.

[125] S. Athar and Z. Wang, "A comprehensive performance evaluation of image quality assessment algorithms," *IEEE Access*, vol. 7, pp. 140030–140070, 2019.

[126] K. Okarma, "Hybrid Feature Similarity Approach to Full-Reference Image Quality Assessment," in *Proc. Int. Conf. Comput. Vis. Graph*, pp. 212–219, Sept. 2012.

[127] K. Okarma, "Combined Image Similarity Index," *Opt. Rev.*, vol. 19, pp. 349–354, Sept. 2012.

[128] K. Okarma, "Quality assessment of images with multiple distortions using combined metrics," *Elektronika ir Elektrotechnika*, 2014.

[129] K. Okarma, P. Lech, and V. V. Lukin, "Combined full-reference image quality metrics for objective assessment of multiply distorted images," *Electronics*, vol. 10, no. 18, 2021.

[130] O. I. Ieremeiev, V. V. Lukin, N. N. Ponomarenko, K. O. Egiazarian, and J. Astola, "Combined full-reference image visual quality metrics," *Electron. Imag.*, vol. 2016, pp. IPAS–180:1–10, Feb. 2016.

[131] K. Okarma, "Combined Full-Reference Image Quality Metric Linearly Correlated with Subjective Assessment," in *Proc. Int. Conf. AI. Soft Comput.*, pp. 539–546, June 2010.

[132] K. Okarma, "Extended Hybrid Image Similarity–Combined Full-Reference Image Quality Metric Linearly Correlated with Subjective Scores," *Elektronika ir Elektrotechnika*, 2013.

[133] L. Jin, K. Egiazarian, and C.-C. J. Kuo, "Perceptual image quality assessment using block-based multi-metric fusion (BMMF)," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 1145–1148, Mar. 2012.

[134] T. Liu and W. Lin and C.-C. J. Kuo, "Image Quality Assessment Using Multi-Method Fusion," *IEEE Trans. Image Process.*, vol. 22, pp. 1793–1807, May 2013.

[135] A. Chetouani, "An Image Quality Metric with Reference for Multiply Distorted Image," in *Proc. Int. Conf. Adv. Concepts Intell. Vis. Syst.*, pp. 477–485, Oct. 2016.

[136] L. Zhang, L. Zhang, and X. Mou, "RFSIM: A feature based image quality assessment metric using Riesz transforms," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, pp. 321–324, Sept. 2010.

[137] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, J. Astola, M. Carli, and F. Battisti, "TID2008–A Database for Evaluation of Full-Reference Visual Quality Assessment Metrics," *Adv. Modern Radioelectron.*, vol. 10, no. 4, pp. 30–45, 2009.

[138] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image Quality Assessment Based on a Degradation Model," *IEEE Trans. Image Process.*, vol. 9, pp. 636–650, Apr. 2000.

[139] D. M. Chandler and S. S. Hemami, "VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images," *IEEE Trans. Image Process.*, vol. 16, pp. 2284–2298, Sept. 2007.

[140] A. Mansouri, A. M. Aznaveh, F. Torkamani-Azar, and J. A. Jahanshahi, "Image quality assessment using the singular value decomposition theorem," *Optical Review*, vol. 16, no. 2, pp. 49–53, 2009.

[141] T. Liu, W. Lin, and C.-C. J. Kuo, "A multi-metric fusion approach to visual quality assessment," in *Proc. Int. Workshop Qual. Multimedia Exper. (QoMEX)*, pp. 72–77, Sept. 2011.

[142] V. V. Lukin, N. N. Ponomarenko, O. I. Ieremeiev, K. O. Egiazarian, and J. Astola, "Combining full-reference image visual quality metrics by neural network," in *Proc. SPIE Electron. Imag.*, vol. 9394, pp. 93940K:1–93940K:12, Mar. 2015.

[143] P. Ye, J. Kumar, and D. Doermann, "Beyond Human Opinion Scores: Blind Image Quality Assessment based on Synthetic Scores," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014.

[144] G. V. Cormack, C. L. A. Clarke, and S. Buettcher, "Reciprocal Rank Fusion outperforms Condorcet and Individual Rank Learning Methods," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2009.

[145] L. Akritidis, A. Fevgas, and P. Bozanis, "An iterative distance-based model for unsupervised weighted rank aggregation," in *2019 International Conference on Web Intelligence*, pp. 358–362, 2019.

[146] L. Akritidis, A. Fevgas, P. Bozanis, and Y. Manolopoulos, "An unsupervised distance-based model for weighted rank aggregation with list pruning," *Expert Syst. Appl.*, vol. 202, sep 2022.

[147] A. Klementiev, D. Roth, and K. Small, "Unsupervised rank aggregation with distance-based models," in *Proceedings of the 25th International Conference on Machine Learning*, p. 472–479, Association for Computing Machinery, 2008.

[148] A. Klementiev, D. Roth, K. Small, and I. Titov, "Unsupervised rank aggregation with domain-specific expertise," in *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, 2009.

[149] A. Klementiev, D. Roth, and K. Small, "An unsupervised learning algorithm for rank aggregation," in *Machine Learning: ECML 2007* (J. N. Kok, J. Koronacki, R. L. d. Mantaras, S. Matwin, D. Mladenič, and A. Skowron, eds.), pp. 616–623, Springer Berlin Heidelberg, 2007.

[150] C. L. MALLOWS, "NON-NULL RANKING MODELS. I," *Biometrika*, vol. 44, pp. 114–130, 06 1957.

[151] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "Dipiq: Blind image quality assessment by learning-to-rank discriminable image pairs," *Trans. Img. Proc.*, vol. 26, p. 3951–3964, aug 2017.

[152] K. Ma, X. Liu, Y. Fang, and E. P. Simoncelli, "Blind image quality assessment by learning from multiple annotators," in *2019 IEEE International Conference on Image Processing*, pp. 2344–2348, 2019.

[153] D. He, D. Cai, J. Zhou, J. Luo, and S.-L. Chen, "Restoration of out-of-focus fluorescence microscopy images using learning-based depth-variant deconvolution," *IEEE Photonics Journal*, vol. 12, no. 2, pp. 1–13, 2020.

[154] H. Zhao, Z. Ke, N. Chen, S. Wang, K. Li, L. Wang, X. Gong, W. Zheng, L. Song, Z. Liu, *et al.*, "A new deep learning method for image deblurring in optical microscopic systems," *Journal of biophotonics*, vol. 13, no. 3, p. e201960147, 2020.

[155] C. Jiang, J. Liao, P. Dong, Z. Ma, D. Cai, G. Zheng, Y. Liu, H. Bu, and J. Yao, "Blind deblurring for microscopic pathology images using deep learning networks," *arXiv preprint arXiv:2011.11879*, 2020.

[156] C. Zhang, H. Jiang, W. Liu, J. Li, S. Tang, M. Juhas, and Y. Zhang, "Correction of out-of-focus microscopic images by deep learning," *Computational and Structural Biotechnology Journal*, vol. 20, pp. 1957–1966, 2022.

[157] J. Wang and B. Han, "Defocus deblur microscopy via head-to-tail cross-scale fusion," in *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 2081–2086, IEEE, 2022.

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

[159] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[160] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241, Springer, 2015.

[161] D. Nicmanis, "Deep learning based focus interpolation for whole slide images," master's thesis, Uppsala University, 2022.

[162] 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*, vol. 7, pp. 124–136, 2020.

[163] I. Mazilu, S. Wang, S. Dummer, R. Veldhuis, C. Brune, and N. Strisciuglio, "Defocus blur synthesis and deblurring via interpolation and extrapolation in latent space," in *Computer Analysis of Images and Patterns*, (Cham), pp. 201–211, Springer Nature Switzerland, 2023.

[164] Y. Leng, *Materials characterization: introduction to microscopic and spectroscopic methods*. John Wiley & Sons, 2013.

[165] R. Brixtel, S. Bougleux, O. Lézoray, Y. Caillot, B. Lemoine, M. Fontaine, D. Nebati, and A. Renouf, "Whole slide image quality in digital pathology: Review and perspectives," *IEEE Access*, vol. 10, pp. 131005–131035, 2022.

[166] S. W. Jahn, M. Plass, and F. Moinfar, "Digital pathology: advantages, limitations and emerging perspectives," *Journal of clinical medicine*, vol. 9, no. 11, p. 3697, 2020.

[167] Q. Li, X. Liu, K. Han, C. Guo, J. Jiang, X. Ji, and X. Wu, "Learning to autofocus in whole slide imaging via physics-guided deep cascade networks," *Optics Express*, vol. 30, no. 9, pp. 14319–14340, 2022.

[168] X. Geng, X. Liu, S. Cheng, and S. Zeng, "Cervical cytopathology image refocusing via multi-scale attention features and domain normalization," *Medical Image Analysis*, vol. 81, p. 102566, 2022.

[169] H. Pinkard, Z. Phillips, A. Babakhani, D. A. Fletcher, and L. Waller, "Deep learning for single-shot autofocus microscopy," *Optica*, vol. 6, no. 6, pp. 794–797, 2019.

[170] C. McQuin, A. Goodman, V. Chernyshev, L. Kamentsky, B. A. Cimini, K. W. Karhohs, M. Doan, L. Ding, S. M. Rafelski, D. Thirstrup, *et al.*, "Cellprofiler 3.0: Next-generation image processing for biology," *PLoS biology*, vol. 16, no. 7, 2018.

[171] A. Janowczyk, R. Zuo, H. Gilmore, M. Feldman, and A. Madabhushi, "Histoqc: an open-source quality control tool for digital pathology slides," *JCO clinical cancer informatics*, vol. 3, pp. 1–7, 2019.

[172] A. Gupta, P. J. Harrison, H. Wieslander, N. Pielawski, K. Kartasalo, G. Partel, L. Solorzano, A. Suveer, A. H. Klemm, O. Spjuth, *et al.*, "Deep learning in image cytometry: a review," *Cytometry Part A*, vol. 95, no. 4, pp. 366–380, 2019.

[173] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature medicine*, vol. 25, no. 1, pp. 44–56, 2019.

[174] Z. Wang, S. Athar, and Z. Wang, "Blind quality assessment of multiply distorted images using deep neural networks," in *International Conference on Image Analysis and Recognition*, pp. 89–101, 2019.

[175] B. T. Atmaja and M. Akagi, "Evaluation of error- and correlation-based loss functions for multitask learning dimensional speech emotion recognition," *Journal of Physics: Conference Series*, vol. 1896, 2020.

[176] Z. Wang, M. Hosseini, A. Miles, K. Plataniotis, and Z. Wang, "Focuslitenn: High efficiency focus quality assessment for digital pathology," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, Springer International Publishing, 2020.

[177] K. Stacke, G. Eilertsen, J. Unger, and C. Lundström, "A closer look at domain shift for deep learning in histopathology," *arXiv preprint arXiv:1909.11575*, 2019.

[178] S. Athar, Z. Wang, and Z. Wang, "Deep neural networks for blind image quality assessment: addressing the data challenge," *arXiv preprint arXiv:2109.12161*, 2021.

[179] C. C. Yang and S. H. Kwok, "Efficient gamut clipping for color image processing using LHS and YIQ," *Opt. Eng.*, pp. 701–711, 2003.

[180] H. Chang, H. Yang, Y. Gan, and M. Wang, "Sparse Feature Fidelity for Perceptual Image Quality Assessment," *IEEE Trans. Image Process.*, vol. 22, pp. 4007–4018, Oct. 2013.

[181] A. Balanov, A. Schwartz, Y. Moshe, and N. Peleg, "Image quality assessment based on DCT subband similarity," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, pp. 2105–2109, Sept. 2015.

[182] L. Li, H. Cai, Y. Zhang, W. Lin, A. C. Kot, and X. Sun, "Sparse Representation-Based Image Quality Index With Adaptive Sub-Dictionaries," *IEEE Trans. Image Process.*, vol. 25, pp. 3775–3786, Aug. 2016.

[183] T. Wang, L. Zhang, H. Jia, B. Li, and H. Shu, "Multiscale contrast similarity deviation: An effective and efficient index for perceptual image quality assessment," *Signal Process.: Image Commun.*, 2016.

[184] I. Lissner, J. Preiss, P. Urban, M. S. Lichtenauer, and P. Zolliker, "Image-Difference Prediction: From Grayscale to Color," *IEEE Trans. Image Process.*, vol. 22, pp. 435–446, Feb. 2013.

[185] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient Magnitude Similarity Deviation: A Highly Efficient Perceptual Image Quality Index," *IEEE Trans. Image Process.*, pp. 684–695, 2014.

[186] S. Rezazadeh and S. Coulombe, "Low-complexity computation of visual information fidelity in the discrete wavelet domain," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2010.

[187] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, "Blind Image Quality Assessment Based on High Order Statistics Aggregation," *IEEE Trans. Image Process.*, vol. 25, pp. 4444–4457, Sept. 2016.

[188] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, "End-to-End Blind Image Quality Assessment Using Deep Neural Networks," *IEEE Trans. Image Process.*, vol. 27, pp. 1202–1213, Mar. 2018.

[189] W. Xue, L. Zhang, and X. Mou, "Learning without Human Scores for Blind Image Quality Assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 995–1002, June 2013.

[190] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, "LIVE Image Quality Assessment Database Release 2." Available: http://live.ece.utexas.edu/research/Quality/subjective.htm.

[191] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. J. Kuo, "Image database TID2013: Peculiarities, results and perspectives," *Signal Process.: Image Commun.*, vol. 30, pp. 57–77, Jan. 2015.

[192] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, pp. 011006:1–011006:21, Jan. 2010.

[193] X. Liu, M. Pedersen, and J. Y. Hardeberg, "CID:IQ – A New Image Quality Database," in *Proc. Int. Conf. Image, Signal Process. (ICISP)*, pp. 193–202, July 2014.

[194] W. Sun, F. Zhou, and Q. Liao, "MDID: A multiply distorted image database for image quality assessment," *Pattern Recognit.*, vol. 61, pp. 153–168, Jan. 2017.

[195] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Hybrid No-Reference Quality Metric for Singly and Multiply Distorted Images," *IEEE Trans. Broadcast.*, vol. 60, pp. 555–567, Sept. 2014.

[196] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik, "Objective quality assessment of multiply distorted images," in *Proc. Asilomar Conf. Signals, Syst., Comput. (ASILOMAR)*, pp. 1693–1697, Nov. 2012.

[197] S. Corchs and F. Gasparini, "A Multidistortion Database for Image Quality," in *Proc. Int. Workshop Comput. Color Imag.*, 2017.

[198] Z. Duanmu, W. Liu, Z. Wang, and Z. Wang, "Quantifying visual image quality: A bayesian view," *Annual Review of Vision Science*, 2021.

[199] R. Lott, J. Tunnicliffe, E. Sheppard, J. Santiago, C. Hladik, M. Nasim, K. Zeitner, T. Haas, S. Kohl, and S. Movahedi-Lankarani, "Practical guide to specimen handling in surgical pathology," *College of American Pathologists*, pp. 1–52, 2015.

[200] Y. Zhang, P. Zheng, W. Yan, C. Fang, and S. S. Cheng, "A unified framework for microscopy defocus deblur with multi-pyramid transformer and contrastive learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2024.

[201] Y. Zhou, H. Wang, Y. Bai, Y. Wan, C. Jin, M. Chen, and X. Teng, "Digital pathology image deblurring via local focus quality assessment," in *ICASSP 2024-2024 IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2165–2169, IEEE, 2024.

[202] 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*, vol. 16, no. 12, pp. 1323–1331, 2019.

[203] W. Zhang and W.-K. Cham, "Single image focus editing," in *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pp. 1947–1954, IEEE, 2009.

[204] P. Sakurikar, I. Mehta, V. N. Balasubramanian, and P. Narayanan, "Refocusgan: Scene refocusing using a single image," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 497–512, 2018.

[205] B. Busam, M. Hog, S. McDonagh, and G. Slabaugh, "Sterefo: Efficient image refocusing with stereo vision," in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pp. 0–0, 2019.

[206] N. Dey, A. Boucher, and M. Thonnat, "Image formation model of a 3d translucent object observed in light microscopy," in *Proceedings. International Conference on Image Processing*, vol. 2, pp. II–II, 2002.

[207] A. K. Khitrin, J. C. Petruccelli, and M. A. Model, "Bright-field microscopy of transparent objects: A ray tracing approach," *Microscopy and Microanalysis*, vol. 23, no. 6, pp. 1116–1120, 2017.

[208] S. Yoo, P. Ruiz, X. Huang, K. He, X. Wang, I. Gdor, A. Selewa, M. Daddysman, N. J. Ferrier, M. Hereld, N. Scherer, O. Cossairt, and A. K. Katsaggelos, "Bayesian approach for automatic joint parameter estimation in 3d image reconstruction from multi-focus microscope," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 3583–3587, 2018.

[209] T. Yamaguchi, H. Nagahara, K. Morooka, Y. Nakashima, Y. Uranishi, S. Miyauchi, and R. Kurazume, "3d image reconstruction from multi-focus microscopic images," in

Image and Video Technology, (Cham), pp. 73–85, Springer International Publishing, 2020.

[210] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[211] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 248–255, June 2009.

[212] R. Zhou, S. Athar, Z. Wang, and Z. Wang, "Deep image debanding," in *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 1951–1955, 2022.

[213] Z. Wang, Z. Chen, and F. Wu, "Thermal to visible facial image translation using generative adversarial networks," *IEEE Signal Processing Letters*, vol. 25, no. 8, pp. 1161–1165, 2018.

[214] Z. Chen, C. Niu, Q. Gao, G. Wang, and H. Shan, "Lit-former: Linking in-plane and through-plane transformers for simultaneous ct image denoising and deblurring," *IEEE Transactions on Medical Imaging*, 2024.

[215] X. Wang, K. Yu, C. Dong, and C. C. Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 606–615, 2018.

[216] L. Wang, Y. Wang, X. Dong, Q. Xu, J. Yang, W. An, and Y. Guo, "Unsupervised degradation representation learning for blind super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10581–10590, 2021.

[217] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 2794–2802, 2017.

[218] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations* (Y. Bengio and Y. LeCun, eds.), 2015.

[219] L. Ruan, B. Chen, J. Li, and M. Lam, "Learning to deblur using light field generated and real defocus images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16304–16313, 2022.

[220] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5728–5739, 2022.

[221] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Multi-stage progressive image restoration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14821–14831, 2021.