Attention Mechanisms for Fine-grained Biomedical Image Analysis



Mengran Fan

Trinity College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2020

Acknowledgements

First of all, I would like to sincerely acknowledge my primary supervisor Prof. Jens Rittscher and co-supervisor Dr. Tapabrata (Rohan) Chakraborty, for their invaluable trust, support, and encouragement throughout my PhD, for guiding me in exploring the breadth and depth of my knowledge throughout the research process, and, importantly, for always being there to support me in pursuing my own ideas and research interests.

I would like to gratefully thank my labmates in the Rittscher research group, for all the inspiring discussions, for the enjoyable and memorable journey we had working together, and for the stimulating and supportive research environment. I would also like to thank my clinical collaborators Dr. Chandan Seth Nanda, Zhihang Yang and Prof. Holm Uhlig, who provided me with valuable clinical imaging data and medical knowledge that enabled the work on this thesis to proceed.

I would also like to acknowledge the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre. Thanks for providing computing resources for the work in this thesis.

Last but not least, I would like to express my sincere gratitude to my beloved parents for their selfless love and generous financial support throughout my education journey. After many years away from home, I hope their original intentions and expectations have not been disappointed. I would also like to thank my boyfriend and best friends for their companionship and encouragement.

Without their constant support, I could not have gone through these four years with no regrets. Thank you very much to everyone I have met along the way. Hope this thesis is just the start of my research journey.

Abstract

Problem: Recently, deep convolutional neural networks have greatly improved our ability to develop robust image analysis methods, especially those tailored to biomedical scenarios. Despite important advances in computer vision, it is usually challenging to achieve the same success on specific biomedical datasets due to certain specific challenges, such as limited size datasets, expensive annotations, high-resolution information, multiple discernible biomedical objects, and lack of interpretability. Therefore, it is still in high demand to investigate automatic image analysis methods with high interpretability and accuracy to transform high-resolution biomedical image data into meaningful quantitative information with weak annotations and limited availability of training data.

Methodology: Considering the challenges of biomedical datasets and the corresponding technical limitations, this thesis aims to incorporate global context-aware information to enable automated image analysis systems to provide reliable diagnostics in clinical applications. The thesis first introduces a hybrid method equipped with U-Net and a global probabilistic model for the segmentation of dense cell populations. This model jointly exploits shape priors and global cellular interactions to compensate for the limitations of standard deep networks. Next, this thesis focuses on attention-based, interpretable fine-grained classification methods for biomedical images that is capable of jointly capturing local high-resolution detail and global background information.

Outcomes: In this thesis, a number of fine-grained feature learning methods were proposed for high-resolution biomedical images. The original contributions are as follows: (1) A global probabilistic model for dense cell nucleus segmentation. (2) A gated attention mechanism for simultaneous localisation of multiple discriminative instances. (3) A multi-task learning scheme for fine-grained feature learning. (4) Two new clinical datasets for biomedical image classification.

Contents

List of Figures **vii**

List of Abbreviations **xv**

1 Introduction **1**
1.1 Introduction . . . 1
1.1.1 Complementary Fine-grained Feature Learning . . . 3
1.1.2 Automatic Analysis for Small High-resolution Biomedical Datasets . . . 4
1.2 Thesis Structure and Contributions . . . 7
1.3 Publications . . . 10

2 Background **11**
2.1 Cell Segmentation . . . 12
2.1.1 Data-driven Methods . . . 12
2.1.2 Model-driven Methods . . . 18
2.1.3 Deep learning-based Methods . . . 20
2.2 Biomedical Image Classification . . . 25
2.2.1 Image Classification in Computer Vision . . . 25
2.2.2 One-stage Deep Classification of Biomedical Images . . . 27
2.2.3 Two-stage High-resolution Classification of Biomedical Images 30
2.3 Relevant Advances in Deep Learning . . . 35
2.3.1 Attention Mechanisms . . . 35
2.3.2 Fine-grained Image Classification . . . 44
2.4 Summary . . . 49

3 Global Probabilistic Models for Cell Segmentation **50**
3.1 Introduction . . . 51
3.2 Deep learning-based Segmentation . . . 54
3.2.1 U-Net Architecture . . . 55
3.2.2 Training Details . . . 57
3.2.3 Experimental Results . . . 58
3.2.4 Failure Cases . . . 59

3.3 Model-based Segmentation 60
3.3.1 Model Construction 61
3.3.2 Algorithm Overview 65
3.3.3 Affinity Functions. 70
3.3.4 Analysis of Segmentation Result 72
3.3.5 Analysis of Global Optimisation 75
3.4 Overall Segmentation 75
3.4.1 Hybrid Segmentation Framework 75
3.4.2 Experimental Results 79
3.5 Conclusions 80

4 Fine-grained Fungi Species Classification through Deep Attention 84
4.1 Introduction 85
4.2 Method 90
4.2.1 Attention Mechanism 90
4.2.2 Feature fusion 93
4.2.3 Multi-task Loss Function 97
4.3 New Clinical Fungi Dataset 98
4.3.1 Clinical background 99
4.3.2 Morphological Features 100
4.3.3 Data Acquisition 103
4.3.4 Dataset Description 104
4.4 Experiments and Results 105
4.4.1 Implementation details 105
4.4.2 Quantitative Comparison with Competing Methods 106
4.4.3 Comparative Evaluation of Attention Mechanisms 109
4.4.4 Visualisation of the Attention Map 110
4.5 Conclusion 113

5 An Interpretable Gated Attention Mechanism 115
5.1 Introduction 116
5.2 Method 119
5.2.1 First Extension: Gated Attention Mechanism 119
5.2.2 Second Extension: Weighted Feature Fusion Scheme 121
5.2.3 Third Extension: Instance Representation Loss 122
5.3 Datasets 124
5.3.1 Clinical Fungi Dataset 124
5.3.2 BreAst Cancer Histology images (BACH) 125
5.3.3 ISIC-2017 Dataset 125
5.3.4 Coeliac Disease Dataset 127

5.4 Experiments and Results 131
5.4.1 Implementation details 131
5.4.2 Evaluation and Analysis on Clinical Fungi Dataset 131
5.4.3 Evaluation and Analysis on BACH dataset 139
5.4.4 Evaluation and analysis on ISIC-2017 Dataset 139
5.4.5 Evaluation and Analysis on Coeliac Disease Dataset 142
5.5 Clinical Background Alignment 144
5.5.1 Breast Cancer Classification from Histopathology Images . . 145
5.5.2 Skin Cancer Classification from Dermatoscopic Images . . . 147
5.5.3 Fungal Species Classification from Mycology Images 148
5.5.4 Coeliac Disease Classification from Histopathological Images 150
5.6 Conclusion 151

6 Conclusions and Future Work 153
6.1 Summary of Contributions 153
6.2 Directions for Future Work 156
6.2.1 Region-based whole-slide image annotation strategy 156
6.2.2 Weakly Supervised Whole-slide Image Classification 159
6.3 Summary 162

References 163

List of Figures

1.1 **The local connectivity in a CNN-based architecture.** In these convolutional layers, each neuron is only connected to a 3×3 pixels in the previous layer. The area labelled in grey shows the receptive field of one neuron in Layer3, and the area labelled in black marks the receptive field of one neuron in Layer 2. 4

1.2 **Examples of breast histology images which highlight the importance of simultaneously capturing high-resolution cellular features and global contextual information.** Methods that can identify these three categories require high-resolution information (red box) to distinguish benign lesions from cancers, and a larger context to distinguish in situ lesions from invasive lesions. . . 5

1.3 **Illustration of Challenges associated with biomedical image analysis.** (a) Multiple cells (yellow box) are densely packed with intensity inhomogeneities due to the diversity of imaging parameters and cell interaction. (b) The existence of artefacts and background impurities (red box) poses challenges to object detection. (c) The images from the same class (benign) show different visual characteristics, but the images from different class exhibit similar appearances. Hence the large intra-class and small inter-class variations may make diagnosis process ambiguous. 6

2.1 Common oversegmentation errors encountered with the watershed segmentation. 15

2.2 Sample result of the SLIC superpixel method. 16

2.3 The overall structure of a fully convolutional network (FCN) [52]. . 21

2.4 **The example experimental results of U-Net trained using an adaptive weighting scheme.** When the weight of edge pixels was assigned to 1, the network identified a cluster of cells as an object. When the boundary weight was 2, most of the pixels inside and the cytoplasm near the boundaries were also predicted as boundaries. . 23

2.5 Architecture of AlexNet[50], VGG Network[51], GooLeNet[70] and ResNet[71]. The figures are taken from the original papers. 26

2.6 **The overall structure for two-stage classification of biomedical images.** CNN cannot be applied to very high-resolution images due to memory limitations and high computational costs. Therefore, a two-stage learning strategy is widely used for high-resolution biomedical image classification. In the first stage, the high-resolution images are partitioned into locally equal-sized patches, which can produce an extremely large number of small patches to train the patch-level CNNs. In the second stage, the patch-level classification results are aggregated into image-level predictions. 31

2.7 **The overall structure for two-stage attention-based classification of biomedical images.** In this paradigm, attention mechanisms are used to select informative patches in high-resolution images. In contrast to the standard two-stage framework, which produces patch-level classification results, this attention-based scheme highlights discriminative patches in the classification task. 42

2.8 **Fine-grained classification VS. general image classification.** The blue box contains images of different species of birds, namely Scott Oriole and Orchard Oriole. The red box gives an images of a dog. Fine-grained recognition focuses on objects belonging to the same basic class, whereas general image classification distinguishes between basic-level categories. 44

3.1 **The standard image acquisition process for multi-channel fluorescent imaging.** 52

3.2 **Illustration of the challenges in automatic segmentation of cell nuclei.** 53

3.3 **U-Net architecture with three stages and 128 channels in the first layer.** 55

3.4 **Failure modes of U-Net based segmentation.** The first row presents the result of U-Net-A which is trained using an equally weighted loss function. It can be observed that the network identifies a clump of cells as one object. The second row shows the failure modes of U-Net-B with a weighted loss function. Although it can separate the clustered nuclei successfully, the delineation of the object boundaries is not very accurate. 60

3.5 **Model-based segmentation.** The figure shows selected results of our model-based segmentation methods that correspond to failure modes that were highlighted in Figure 3.4. 61

3.6 **Illustration of the proposed model-based approach.** The core principle of the proposed segmentation model is to assign image observations Z to a set of object hypotheses H that are associated with model parameters X. The set of image observations Z consists of various region-based image features based on an initial superpixel step. As an initial estimate, each collected superpixel segment is considered as a single cell hypothesis h_i. With the local image features and initial hypotheses, the assignment vector V is directly defined to associate each image segment z_i with a particular hypothesis h_k. It is worth noting that the assignment vector V is treated as a hidden variable since the assignment of segment z_i to hypothesis h_k can not be observed directly. The segmentation on the bottom shows the maximum a posteriori (MAP) estimate of the model parameters X^*. 62

3.7 **Sample result of the SLIC superpixel method.** The SLIC algorithm [27] is widely used to efficiently generate locally consistent patches. It is clear that the over-segmented regions adhere well to the boundaries of the cell nuclei. 67

3.8 **Preliminary result for one sample image.** (a) Input image. (b) Ground truth. (c) Over-segmented result generated by SLIC superpixel algorithm. (d) Resulting edge map of canny edge detector. (e) Segmentation result of the proposed model. 69

3.9 Illumination of edge affinity calculation 72

3.10 **Illustration of typical failure modes in the segmentation results.** In this failure analysis, there are three different types of errors. The images on the left present groundtruth labels, while the images on the right are the segmentation results generated by the proposed segmentation model. (a) Example of under-segmentation error, which describes the situations where a region is not split into the correct number of cell nuclei. (b) Example of over-segmentation error, which is defined as excessive splitting. (c) Example of boundary error. In addition to the two standard types of errors, we also looked for boundary errors, which occur when the segmentation framework does not delineate the correct boundary between a pair of touching nuclei. For instance, a detected nuclear candidate may include a portion of its neighboring nucleus, while the neighboring nucleus is not detected at all. 74

3.11 **Plots of probability variance.** (a) probability of segment 1 belonging to hypothesis 1, (b) probability of segment 1 belonging to hypothesis 2, (c) probability of segment 2 belonging to hypothesis 1, and (d) probability of segment 2 belonging to hypothesis 2 during the iterative process. 76

3.12 **Model visualisation with respect to a small portion of the sample image.** Model visualisation with respect to a small portion of the sample image. (a)-(f) describe the result of 0^{th},1^{st},3^{rd},5^{th},7^{th},9^{th},12^{th} and 15^{th} iterations respectively. The red boundaries indicate valid nuclei hypotheses, while the yellow boundaries delineate the regions of cytoplasm. 77

3.13 **Representative segmentation results using the proposed probabilistic model.** The original images and segmentation results are shown in left and right columns, respectively. 78

3.14 **Overall segmentation algorithm.** Given the segmentation results of U-Net, very basic constraints on shape and size are being used to detect the outliers (failure cases) of U-Net. Detected failure cases are handled explicitly by the proposed global probability model described in the chapter, while the successful segmentation results generated by U-Net are passed directly to the final segmentation map. 80

3.15 **Segmentation performance of the standard U-Net and the proposed hybrid framework.** The input images area shown in the first column. The second column presents the segmentation results of standard U-Net, and the third column shows the segmentation results of the proposed hybrid segmentation approach. 81

4.1 **Classification challenges in biomedical imaging.** Compared to large-scale datasets on natural images (a), well-organised biomedical datasets (b) tend to be small and require expensive human expert annotations. Secondly, while we typically expect one centre-aligned instance (red box) in natural images, there are often multiple discriminative instances (red boxes) in biomedical images, which poses another challenge for feature learning. Finally, high-resolution information and interpretability need to be considered when developing a reliable medical image analysis system. 87

4.2 **An overview of the proposed fine-grained classification framework.** Using only image-level labels, the proposed model learns to localise multiple discriminative instance simultaneously and to integrate the global structural information with local high-resolution features for fine-grained classification. 89

4.3 **Framework for the proposed multiple instance fine-grained classification pipeline.** The network consists of two main modules: (i) a novel attention mechanism for discriminative instances localisation; and (ii) a feature fusion scheme that consolidates the global and local features to facilitate the final predictions. 91

4.4 **Self-attention module.** Given a feature vector $\mathbf{X} \in \mathcal{R}^{h \times w \times c}$, where h, w, c stand for the height, width and channels respectively, the self-attention module XX^TX is used to attend to the entire region of the biomedical object. 93
4.5 **The mechanism of the attention-based soft sampling.** (a) is the original full resolution image and (b) is the attention map generated by the proposed attention module. (c1) and (c2) are the marginal distributions over x and y axis, respectively. (d1) and (d2) are the cumulative distribution functions of (c1) and (c2). (e) illustrates the final sampled images. 94
4.6 **Examples of the attention-based soft sampling.** 96
4.7 **Failure cases of soft sampling.** It can be seen that soft sampling may result in spatial distortion when it applied to multi-instance cases. To keep the shape information of multiple instances, hard sampling were conducted to generate informative regions. 97
4.8 **A schematic diagram of the structure of Aspergillus species.** A more detailed description of the structure is provided in the text. (The figure was created with `https://BioRender.com`.) 101
4.9 **A schematic diagram of the conidial heads of different Aspergillus species.** 102
4.10 **An overview of the process of data acquisition.** 104
4.11 **Examples for the five classes in the Fungi dataset.** 106
4.12 **Visualisation of global attention maps.** (a) Each row shows a sample image from each class represented in the fungal dataset. Our attention maps suppress features associated with background clutters and focus on the whole objects. 111
4.13 **Failure cases of global attention maps within the test set.** 112

5.1 **Framework for the proposed attention-based fine-grained classification pipeline.** In the attention-based localisation module (Section 5.2.1), multiple instances are highlighted based on global image-level features. Details of the proposed semantic modelling and gated mechanism are depicted in Figure 5.2 and Figure 5.3, respectively. However, the number of extracted instances is fixed across the whole dataset, and when the number of discriminative instances in the image is small, this strategy may face a risk that not all sampled instances are meaningful for the final fine-grained prediction. Therefore, a weighted feature fusion scheme (Section 5.2.2) and an instance representation loss (Section 5.2.3) are introduced to enhance the instance-level feature learning. 118

5.2 **Semantic modelling.** In semantic modelling, two dilated convolution blocks [166] were first applied to the convolutional feature vector $X \in \mathbb{R}^{H\times W\times D}$, where H, W, D stand for the height, width and channels respectively. Then the aggregated map $M \in \mathbb{R}^{H\times W\times 1}$ is obtained by computing the sum of all channels in high-level feature maps: $M = \sum_{k=1}^{D} W^T X$ where W^T is the weights of the dilated blocks. Finally, the high-level semantic map is obtained by applying a spatial softmax layer. 120

5.3 **Gated Mechanism.** To measure the discriminability of each channel, the proposed gated mechanism captures the spatial similarity scores between the convolutional feature vector $X \in \mathbb{R}^{H\times W\times D}$ and the high-level semantic map $S \in \mathbb{R}^{HW\times 1}$, where H, W, D stand for the height, width and channels respectively. The hyperbolic tangent ($tanh$) and ReLU functions are used as gated activation layers to filter out the confounding channels. Finally, the attention map is modelled as a gated average of the outputs of the original convolutional feature maps. 121

5.4 **Line plots of the different loss term in Eq. 5.5 in the joint learning process.** 124

5.5 **Sample images from BACH dataset.** 126

5.6 **Sample images from ISIC-2017 dataset.** 126

5.7 **Sample whole-slide image from coleiac disease dataset.** . . . 128

5.8 **Schematic diagram of marsh classification in coeliac disease.** 129

5.9 **Marsh classification in coeliac disease.** The histology images are selected from our coeliac disease dataset. March 0: normal; March 2: increased intraepithelial lymphocytes with crypt hyperplasia; March 3a: mild villous atrophy; March 3b: marked villous atrophy; March 3c: complete villous atrophy. 132

5.10 **Visualisation of the gated attention module for one sample image.** The top row shows the input image, original convolutional feature map and the corresponding semantic map (defined in Eq.5.1). The bottom row shows the final attention map and two representative channels with the lowest and highest attention. 136

5.11 **Visualisations of the attention mechanism in fungi dataset.**This figure shows five challenging cases, which were misclassified by our earlier baseline method, but were correctly addressed in this network. The attention maps in the third column were generated by the network described in Chapter 4, and the attention maps in the last column were produced by architecture described in this chapter. . 137

5.12 **Visualisations of the failure cases within the test set.** It can be seen that when there are very large instances or clusters of small instances, it is difficult for the model to detect the exact object boundaries. 138
5.13 **Visualisation of the attention maps for coeliac disease dataset.** The first column shows sample images from the coeliac disease dataset, and the second and third columns visualise the results of the last convolutional layer in the ResNet and the attention maps generated in our network, respectively. It can be observed that ResNet tends to focus on the entire tissue regions, whereas our network can explicitly identify the informative villous regions. 143
5.14 **Confusion matrices showing the classification performance on coeliac disease dataset.** 144
5.15 **Clinical alignment on breast cancer dataset.** The first row in (a) shows the sample images with bounding box annotations generated by a pathologist. The corresponding attention maps shown in the second row consistently focus on the clinically relevant features. The localised discriminative region B corresponds to the loose connective tissue. In carcinomas, region C highlights the boundary of the ducts, while D_1, D_2 and D_3 focus on the poorly organised clusters of cells, in which cancerous cells spread into surrounding tissues. (Figure (b) was created with `https://BioRender.com`.) Notice that the bounding boxes on the generated attention maps were placed manually by myself. 145
5.16 **Alignment errors in the generated attention maps and expert annotations.** 147
5.17 **Clinical alignment on ISIC-2017 skin cancer dataset.** The attention maps in benign cases highlight the regular round boundaries, whereas in melanoma cases the network tends to focus on the areas with major irregularities in colour space. (Figure (b) was created with `https://BioRender.com`.) 148
5.18 **Clinical alignment on fungi dataset.**The attention maps of different species consistently focus on the regions of conidial heads, especially the regions of vesicles. (Figure (b) was created with `https://BioRender.com`.) 149
5.19 **Clinical alignment on coeliac disease dataset.** The attention maps consistently highlight the most clearly defined villous structures associated with different degrees of severity: normal or mild samples show a preserved villous structure whereas severe cases highlight the shortened and flattened villi. 150

5.20 **Visualisation of attention maps highlighting crypt hyperplasia in disease samples.** 151

6.1 **Sample patches belonging to the cluster of villi.** Most of the prominent patches in the villous region were effectively identified, however, some irrelevant patches containing artifacts or square outlines also belonged to the villous cluster. 157

6.2 **Three patches selected from the same whole-slide image.** The preprocessing process may lead to a large number of redundant patches in the patch-level dataset. 158

6.3 **A comparison of voting-based and deep learning-based feature aggregation.** In contrast to voting-based methods that aggregate on the patch-level classification results, deep learning-based methods extracts the high-dimensional feature vectors and feeds them as input to a second deep network for image-level prediction . 161

List of Abbreviations

1D, 2D, 3D . . One-, two- or three-dimensional, referring in this thesis to spatial dimensions in an image.

DCNN, CNN (Deep) Convolutional Neural Network, deep network architectures.

FCN . . . . . Fully Convolutional Network, a deep network architecture.

LSTM . . . . . Long-Short-Term Memory, a deep network architecture.

RNN . . . . . Recurrent Neural Network, a deep network architecture.

ResNet . . . . Residual Neural Network, a deep network architecture.

MRI . . . . . Magnetic Resonance Imaging, an imaging modality.

CT . . . . . Computed Tomography, an imaging modality.

CAD . . . . . Computer-Aided Diagnosis.

RoI . . . . . Region of Interest.

SLIC . . . . . Simple Linear Iterative Clustering, a superpixel algorithm.

SVM . . . . . Support vector machine, a machine learning algorithm.

MIL . . . . . Multiple instances learning, a machine learning algorithm.

EM . . . . . Expectation-Maximisation, an iterative method to perform maximum likelihood estimation.

CRF . . . . . Conditional Random Fields, a class of statistical models.

ReLU . . . . . Rectified Linear Units, an activation function in deep networks.

SGD . . . . . Stochastic Gradient Descent, an iterative method for optimising an objective function.

TP, TN, FP, FN True Positive, True Negative, False Positive and False Negative.

CDF . . . . . Cumulative Distribution Function of a distribution over random variables.

MMD . . . . . Maximum Mean Discrepancy, a loss function.

AUC . . . . . Area Under the Curve, a classification evaluation metric.

CD . . . . . Coeliac Disease.

AIDS . . . . . Acquired Immune Deficiency Syndrome

HE . . . . . Hematoxylin and Eosin, a staining technique used in histology.

LPCB . . . . . Lactophenol Cotton Blue, a staining techniques for visualising fungal structures.

SDA . . . . . Sabouraud dextrose agar, a type of agar growth medium.

ISIC . . . . . International Skin Imaging Collaboration.

BACH . . . . . BreAst Cancer Histology.

ABCD rules . Asymmetry, Border irregularity, Colour variegation, and Diameter rules, rules for melanoma diagnosis.

1 Introduction

Contents

1.1 Introduction 1
1.1.1 Complementary Fine-grained Feature Learning 3
1.1.2 Automatic Analysis for Small High-resolution Biomedical Datasets 4
1.2 Thesis Structure and Contributions 7
1.3 Publications 10

1.1 Introduction

The recent years have witnessed the impressive performance of deep learning techniques in a wide range of computer vision tasks. The rise of deep learning methods, especially CNNs (Convolutional Neural Networks), has made it possible to develop end-to-end trainable frameworks without the need for complex feature engineering. More importantly, these methods open the opportunity of learning models that can potentially reliably extract information from complex biomedical images, allowing us to recognise subtle patterns that even experts find hard to disambiguate. As a result, deep learning-based techniques have been widely used in microscopic imaging and computational pathology to enable more efficient workflows

in many clinical applications, such as cell segmentation [1], cancer identification [2], and lymph node detection [3].

However, while it performs well for large, well-annotated datasets, it remains a major challenge to tailor deep learning methods to small biomedical datasets with limited annotations. For example, on large datasets, meaningful fine-grained features can be extracted naturally by deep neural networks, while for small datasets, it is still necessary to carefully design complementary fine-grained feature learning where local high-resolution and global structural features can be extracted simultaneously. In addition, limited training data, expensive annotations, high-resolution information, and the presence of multiple biomedical objects also pose significant challenges to the automatic analysis of biomedical images. Further details on the challenges associated with automatic analysis of small biomedical datasets are presented in Sections 1.1.1 and 1.1.2.

Recently, attention mechanisms have attracted great attention in computer vision and biomedical image analysis. It has been shown that the attention mechanisms can effectively enhance the feature representation of CNNs by highlighting informative regions and discarding irrelevant regions. Therefore, this thesis aims to investigate the attention-based techniques to facilitate fine-grained biomedical image analysis and thus enhance the development of reliable and robust computer-aided diagnostic systems for various biomedical scenarios. The core technical contributions of three research chapters are summarised below:

- A hybrid algorithm for delineation of densely packed cells.
- An attention-based multi-task learning scheme for complementary fine-grained feature learning.
- A interpretable gated attention mechanism for localisation of multiple biomedical objects.

1.1.1 Complementary Fine-grained Feature Learning

In recent years, deep learning-based techniques, in particular deep convolutional neural networks (DCNNs), have achieved remarkable success in the computer vision community and are increasingly gaining attention in the field of biological image analysis. In contrast to traditional methods that rely on hand-crafted features, these DCNN-based methods build and optimise the feature extraction process through "end-to-end" learning for specific tasks, thus providing more flexibility and robustness in practice. However, the design of the deep networks is based on the fact that you have a large dataset of low-resolution images, allowing the network to learn discriminative fine-grained features and to optimise millions of parameters in a large number of iterations. In the case of small high-resolution biomedical images, we will need specific model design to address fine-grained feature learning.

Standard DCNN-based models are usually insufficient to resolve the complementary feature learning where local high-resolution and global structural features can be simultaneously extracted. A convolutional neural network is a specialised kind of feed-forward neural network where each neuron in a convolutional layer is only connected to a certain spatial area in the previous layer and the learnt weights are shared across different spatial regions (Figure 1.1). Compared to the fully-connected manner, this local connectivity allows the network to efficiently exploit meaningful visual patterns in small regions of the input image. However, the available contextual information at each step is relatively local, leading to the restriction in perceiving global knowledge. Some methods [4] address this problem by downscaling the high-resolution input images to expand the receiving field so that the network can take into account the global properties of the entire images, however, a large amount of fine-grained high-resolution information is discarded during this downscaling process. Unlike most natural images, there is a clear need to capture both global structural information and local high-resolution features in biomedical image analysis. For example, Figure 1.2 shows some breast histology images from three categories, benign, in situ carcinoma and invasive carcinoma. It can be observed that the classification of breast histology images is particularly challenging

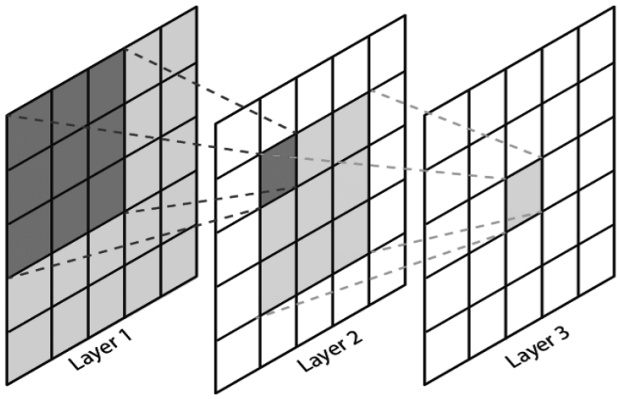


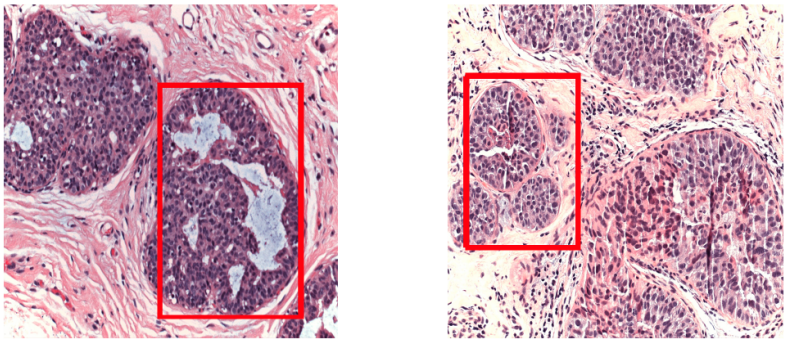
Figure 1.1: The local connectivity in a CNN-based architecture. In these convolutional layers, each neuron is only connected to a 3×3 pixels in the previous layer. The area labelled in grey shows the receptive field of one neuron in Layer3, and the area labelled in black marks the receptive field of one neuron in Layer 2.

since benign lesions and in situ carcinomas have a similar spatial organisation of cells while in situ carcinomas are visually similar to invasive lesions at the cellular level. Thus, methods capable of identifying these three categories need to be able to simultaneously use high-resolution cellular appearance to distinguish benign lesions from in situ carcinoma, and incorporate a large context covering adjacent tissues to distinguish in situ lesions from invasive lesions.

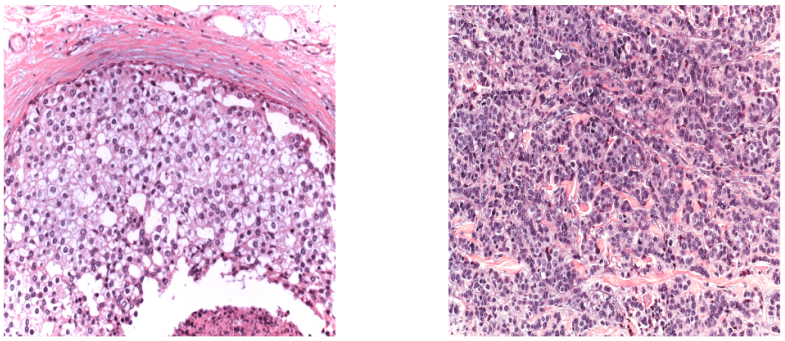
Therefore, automated methods that can overcome the limitations of conventional DCNNs while leveraging both local high-resolution information and global contextual knowledge deserve further exploration, especially for small biomedical datasets. This line of research is also the subject of this thesis, which hopefully will improve the utility of deep networks in real-world applications.

1.1.2 Automatic Analysis for Small High-resolution Biomedical Datasets

Recently, deep convolutional neural networks (DCNNs) have been widely used to achieve state-of-the-art performance for various image analysis tasks. However, DCNNs are generally designed for large, well-organised datasets consisting of large numbers of low-resolution natural images in the computer vision community. In



(a) Benign VS *In situ* Carcinoma



(b) *In situ* Carcinoma VS Invasive carcinoma

Figure 1.2: Examples of breast histology images which highlight the importance of simultaneously capturing high-resolution cellular features and global contextual information. Methods that can identify these three categories require high-resolution information (red box) to distinguish benign lesions from cancers, and a larger context to distinguish in situ lesions from invasive lesions.

biomedical image analysis, there are many real-world environments in which we have access to only limited training data that contain high-resolution information. Unfortunately, in the case of small, high-resolution biomedical datasets, the performance of deep neural networks typically degrades dramatically due to a number of specific technical challenges: (1) Standard DCNNs are prone to overfitting due to the limited amount of training data. (2) Human annotations suffer from issues such as limited reproducibility, domain-specific expertise and unsatisfied consumption of time and labour. (3) Most DCNNs are sufficiently powerful to extract regions of interest from a large and diverse dataset, whereas these models do not have strong attention ability for a small dataset. (4) Due to computational and memory constraints, mainstream

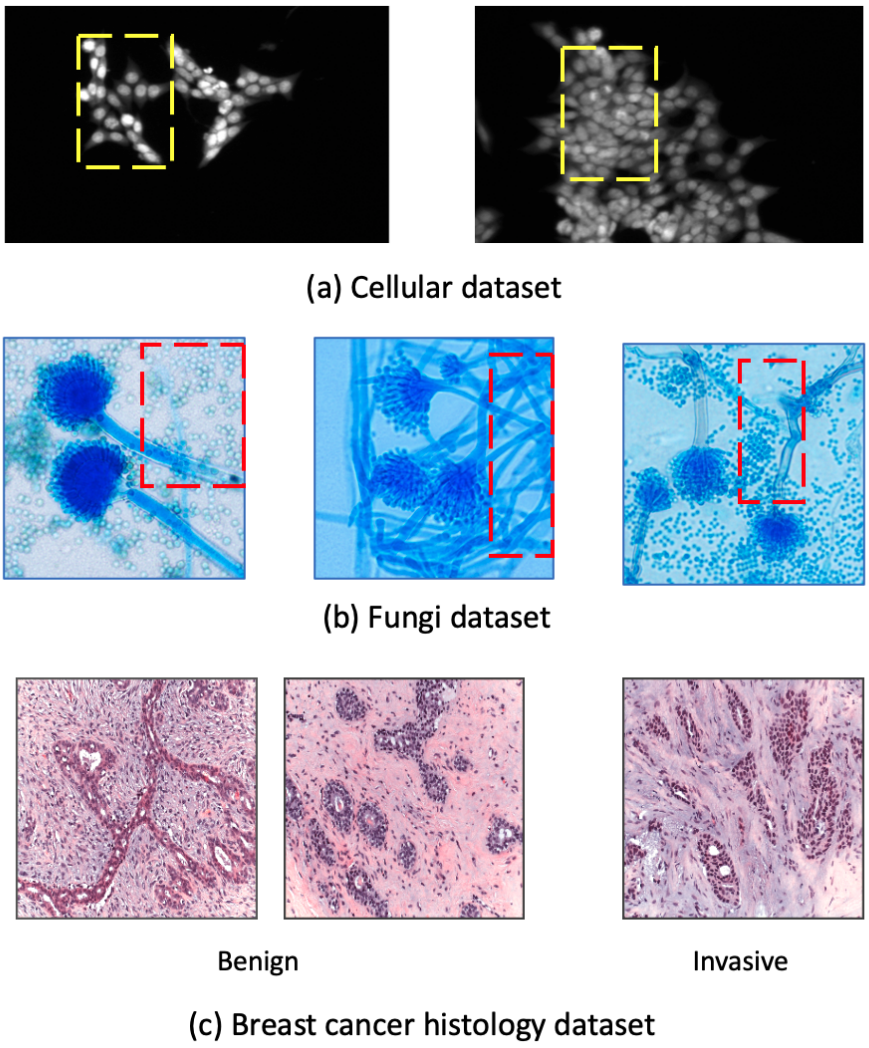


Figure 1.3: Illustration of Challenges associated with biomedical image analysis. (a) Multiple cells (yellow box) are densely packed with intensity inhomogeneities due to the diversity of imaging parameters and cell interaction. (b) The existence of artefacts and background impurities (red box) poses challenges to object detection. (c) The images from the same class (benign) show different visual characteristics, but the images from different class exhibit similar appearances. Hence the large intra-class and small inter-class variations may make diagnosis process ambiguous.

methods applied to high-resolution biomedical images follow a patch-based two-stage approach that divides the high-resolution images into smaller patches that are processed independently by the deep networks for training and inference. While these methods can extract a set of high-resolution patches, they tend to ignore the global long-range correspondences and suppress important fine-grained information, such as the density or shape of biomedical objects.

In addition to technical challenges, there are some general difficulties that are common to most biomedical datasets. Figure 1.3 shows some selected examples from the experimental datasets used in this thesis, illustrating the main challenges. Firstly, the existence of packed structures such as the densely aggregated cells in Figure 1.3a (yellow box) makes robust object segmentation/detection difficult. In addition, the variations in lighting conditions across multiple objects (e.g. the image in the left side of Figure 1.3(a)), as well as impurities and artefacts in the background (e.g. water droplets, bubbles and dust particles), also pose challenges to the detection of main objects. Especially when the amount of training data is limited, the resulting preference may be biased on confounding background clutter. Last but not least, the subtle inter-class variance and a high level of intra-class variation in pose, viewpoint, orientation and scale result in a particular problem in consistent feature learning. For example, two classes of breast cancer histology image are shown in Figure 1.3c. It can be observed that the two images shown on the left are from the same category but show different visual patterns, while the second and third images exhibit similar appearances even though they belong to different categories. These factors are commonly seen in most biomedical datasets and result in difficulties in fine-grained feature learning. Considering the aforementioned challenges of biomedical datasets and the corresponding technical limitations, it is still desirable to investigate automatic image analysis methods to translate the high-resolution biomedical imaging data into meaningful and quantitative information, with limited availability of training data and weak annotations.

1.2 Thesis Structure and Contributions

This thesis presents a number of automatic image analysis methods for a variety of biomedical applications, such as cell segmentation, fungal species recognition, skin cancer identification and histology image analysis. It aims to capture complementary fine-grained features from different small, high-resolution biomedical imaging datasets. Overall, Chapter 1 outlines the limitations of standard DCNNs and the importance of complementary fine-grained feature learning. Chapter 2 presents

a comprehensive literature review. Chapter 3 provides a preliminary study that incorporates global contextual information to compensate for the shortcomings of standard DCNNs. Chapters 4 and 5 delve into the performance of simultaneously capturing local high-resolution and global structural information for fine-grained feature learning. Chapter 6 summarises the main contributions. Further details are as follows:

- **Chapter 1** is the introductory chapter which elaborates challenges and limitations that exist in automatic analysis for small biomedical datasets. This chapter also provides an overview of the main research direction and related contributions.

- **Chapter 2** presents a comprehensive review of existing methods and possible research directions relevant to this thesis. Two related clinical issues and related challenges are presented: cell segmentation and biomedical image classification. This chapters also reviews specific advances in deep learning and computer vision that are particularly relevant to the work presented in this thesis.

- **Chapter 3** examines an approach based on a global optimal model for robust and reliable segmentation of individual objects in dense cell populations. The core of this pipeline is the combination of a DCNN-based segmentation method with a global probabilistic model that explicitly enforces prior shape constraints and global cellular interactions. In this chapter, the U-Net model [5] is used to evaluate the performance of deep learning methods in nucleus segmentation. Our experiments show that deep learning-based methods are not sufficient to locate accurate boundaries for individual cells at the pixel level, especially when we only have access to a limited amount of training data. Although one can overcome this limitation with additional training data or data augmentation techniques, it is still important to explore a robust segmentation framework for automatic nucleus segmentation in fluorescence microscopy

images. This chapter provides preliminary and basic steps for analysing small biomedical datasets and incorporating global contextual information.

- **Chapter 4** turns to another open problem in biomedical image analysis: the fine-grained classification of microscopic imaging data. As opposed to general image recognition, many real-life applications may involve classifying subclasses within the same basic class, e.g., species identification in biomedical image analysis. However, such a task remains challenging for existing methods as it requires the ability to locate and distinguish subtle visual differences in locally discriminative regions within subordinate categories. To this end, this chapter describes a simple yet effective attention-based multi-instance deep architecture which simultaneously integrates local high-resolution and global structural features in an end-to-end fashion. In addition to the development of new algorithms, this chapter also presents a new clinical mycology dataset of fungi images comprising of five species that most commonly affect humans.

- **Chapter 5** showcases the superiority of simultaneously exploring local high-resolution and global structural features for fine-grained classification of microscopic imaging data. This chapter explores three different technical extensions to the work presented in Chapter 4. It is implemented by first using a lightweight gated attention mechanism to detect multiple discriminative instances simultaneously, and then using a weighted feature fusion scheme and an instance representation loss to improve the final image-level classification performance. The proposed network is evaluated on four different biomedical datasets, thus demonstrating that it is a powerful general approach that can be applied to such fine-grained, multi-instance biomedical image classification problems. This chapter also introduces a new biomedical dataset consisting of 392 whole-slide histopathology images of coeliac disease.

- **Chapter 6** summarises the major contributions of this thesis and discusses several limitations of the current work, which further drive the directions of future research.

1.3 Publications

Chapter 3 **Mengran Fan** and Jens Rittscher. Global probabilistic models for enhancing segmentation with convolutional networks. *Proceedings of the International Symposium on Biomedical Imaging (ISBI)*, 2018. - Oral presentation.

Chapter 4 **Mengran Fan**, Tapabrata Chakraborti, Eric I-Chao Chang, Yan Xu, Jens Rittscher. Fine-grained multi-instance classification in microscopy through deep attention. *Proceedings of the International Symposium on Biomedical Imaging (ISBI)*, 2020. - Oral presentation.

Chapter 5 **Mengran Fan**, Tapabrata Chakraborti, Eric I-Chao Chang, Yan Xu, Jens Rittscher. Microscopic fine-grained instance classification through deep attention. *International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI)*, 2020.

2
Background

Contents

2.1 Cell Segmentation 12
2.1.1 Data-driven Methods 12
2.1.2 Model-driven Methods 18
2.1.3 Deep learning-based Methods 20
2.2 Biomedical Image Classification 25
2.2.1 Image Classification in Computer Vision 25
2.2.2 One-stage Deep Classification of Biomedical Images . . 27
2.2.3 Two-stage High-resolution Classification of Biomedical Images 30
2.3 Relevant Advances in Deep Learning 35
2.3.1 Attention Mechanisms 35
2.3.2 Fine-grained Image Classification 44
2.4 Summary 49

This chapter provides a comprehensive review of the literature relevant to this thesis. The order of discussion roughly corresponds to the arrangement of the chapters in this thesis. It starts with an overview of the two specific research questions addressed in this thesis: cell segmentation (Section 2.1) and biomedical image classification (Section 2.2). With regard to each theme, clinical motivations and key challenges are briefly mentioned, followed by a discussion of the relevant scientific literature. Finally, Section 2.3 provides a brief overview of

some recent advances in deep learning and computer vision that have triggered the novelty of this thesis.

2.1 Cell Segmentation

Cell segmentation is widely considered to be the most important and fundamental stage in quantifying cellular mechanisms. With the increasing number of cell images, traditional methods of segmenting individual cells with manual annotation are not feasible. Therefore, reliable and powerful computer-aided techniques are required to process large datasets and extract cellular structures. This section aims to provide a comprehensive review of automated cell segmentation methods. We first discuss two traditional techniques: (1) data-driven algorithms (Section 2.1.1), which use only information extracted from image data to determine the optimal segmentation, and (2) model-driven algorithms (Section 2.1.2), which take into account prior knowledge to support the segmentation of target objects of interest. More advanced deep learning-based approaches are then discussed in Section 2.1.3.

2.1.1 Data-driven Methods

The fundamental difference between data-driven and model-driven methods is whether or not high-level prior information is incorporated into the segmentation process. Data-driven methods start with a single pixel or small over-segmented fragment and iteratively merge regions based on local similarity measurements. These methods, also known as bottom-up segmentation methods, do not require any assumptions or modelling of the underlying objects. However, since the bottom-up approaches only consider pairs of regions at each step, they are inadequate for situations where high ambiguity with respect to local content exists. In contrast, model-driven approaches are guided by some global prior knowledge, including size, shape, or other anatomical morphological patterns. While such top-down segmentations provide the opportunity to consider multiple connected segments simultaneously, they have the limitation of requiring relatively reliable prior knowledge. If the assumptions are incorrect, the corresponding segmentation

model may lead to failure. Thus, both data-driven and model-driven segmentation strategies have their own advantages and disadvantages. In this section, we will focus on three main data-driven segmentation strategies: threshold-based, region-based, and edge-based approaches.

Threshold-based methods. One of the most basic and straightforward segmentation techniques is intensity thresholding [6], which sets a global or local threshold based on a statistical analysis of the intensity histogram. The grayscale intensity of each pixel is compared to the threshold, thus classifying that pixel into objects or background. As a key to these methods, a number of automatic threshold selection algorithms have been proposed to adaptively and iteratively find an optimal threshold value. For example, a popular threshold selection method is called ISODATA thresholding, proposed by Ridler and Calvard [7]. First, a value in the middle of the intensity domain is set as the initial threshold T. Then, the mean A1 of all pixels greater than T and the mean A2 of all pixels less than T are computed. The average of A1 and A2 is used as the new threshold for the next iteration. This operation is repeated until there is no significant change in the threshold value. More complex threshold selection methods can be found in [8][9]. However, the thresholding technique, which is valid only for images with significant differences in grayscale levels, does not work well on cell images, where touching cells, uneven illumination, and uneven intensity are easily observed. In addition, the algorithm treats the segmentation task as a single-pixel classification task, where neither spatial location nor contextual information is considered. Currently, due to its simplicity and speed, it is still widely used as a preprocessing step in pipelines to segment foreground [10][11] from complex backgrounds on cell images. See [12] for a discussion of the threshold method and cell nucleus segmentation.

Region-based methods. Unlike the thresholding methods above, the region-based methods explicitly include spatial information about the pixels in the image. The goal of these methods is to find homogeneous regions in the image. "Regional growth" and "split-and-merge" are two of the best-known techniques in this category.

The main idea of region growth, proposed by Adams and Bischof [13], is to select some seeds as a base and iteratively grow this region by adding unassigned adjacent pixels, thus the image can be segmented into a number of regions containing a starting seed connection. For accurate and reliable segmentation, it is clear that two decisions need to be considered: the selection of the seeds and the setting of the growth criteria. In terms of seed selection, both manual seed marking and automatic seeding have been described in the literature. Adams and Bischof proposed an automatic seed selection scheme for X-ray images of the human chest, but it is not straightforward to transfer this algorithm to other biomedical problems [13]. Fan et al.[14] proposed an algorithm that used the centre point between adjacent colour edges as the initial growth point, where seed selection might lead to image over-segmentation. Shih and Cheng [15] improved Fan et al.'s segmentation algorithm by proposing an automatic seed region growth algorithm for colour image segmentation. However, if the image has highly coloured textures, their algorithm may fail to produce satisfactory results. Considering the large number of cells in biomedical imaging data, it is very difficult to design an automatic selection algorithm to ensure the exact placement of a seed in each nucleus, especially if the cells are tightly packed together or have highly variable intensity profiles. More importantly, the segmentation results are strongly dependent on seed selection, regardless of how we set up the starting seeds. Also, Mehnert and Jackway [16] pointed out another drawback of seeded region growth: the algorithm also strongly relied on the order in which image pixels were processed, and proposed to use parallel processing and re-examination to eliminate the order dependency. Regarding growth criteria, Hojjatoleslami and Kittler [17] used average and peripheral contrast to control the growth of regions. Wan and Higgins [18] proposed a symmetric region growth algorithm, which defined a set of growth criteria to make the algorithm insensitive to initial seeds. Anoraganingrum et al. [19] proposed an adaptive region growth algorithm for the segmentation of melanoma cells (tumour cells) as well as lymphocytes (blood cells) in 3D collagen gels. Instead of using predefined fixed thresholds, they continuously adjust the thresholds based on local intensity

histograms. Extensive work improved this algorithm by incorporating edges, shapes, and other prior knowledge about the region to be grown [20, 21].

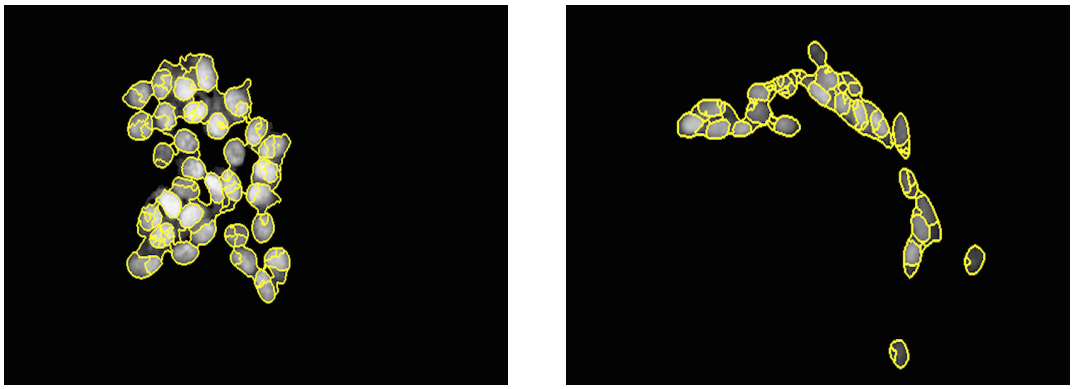


Figure 2.1: Common oversegmentation errors encountered with the watershed segmentation.

The watershed method [22], a nonparametric algorithm that relies on mathematical morphology, is the best known of these algorithms and has been widely used for biomedical image analysis. The main difference between the watershed transform and conventional region growing is that the watershed method works according to intensity layers rather than neighbour layers, which is very effective for segmenting touching cells within the range of intensity variation. However, this approach is also prone to over-segmentation due to the complex inhomogeneous intensities in cellular images. Figure 2.1 shows the segmentation results of the watershed algorithm on our cellular dataset presented in Chapter 3. To address this limitation, some variants of the algorithm have been introduced, but most of them are still very sensitive to the initial selection of markers or require further post-processing steps. Two typical extensions of the original watershed algorithm have been proposed by [23, 24]. The former proposed a rule-based heuristic merging technique to reduce over-segmentation, while the latter created an interactive system that utilised human visual inspection.

On the other hand, split-and-merge is another major strategy of region-based approaches. By definition, the first step is to divide the whole image into a number of smaller patches, and the next step is to merge these separated subregions based on some specified rules. The main advantages compared to the traditional method

of region growth are that there is no need to select initial seed points as inputs and that the required memory is considerably reduced. [25] applied this method to the segmentation of overlapping cell nuclei. Adiga and Chaudhuri [26] proposed a new algorithm to detect the number and location of nuclei in HE-stained tissue samples. They partially solved the speed problem and reduced the execution time in a data-parallel environment. One of the most successful methods is the SLIC algorithm [27] introduced by Achanta et al. They adapted k-means clustering to generate superpixels that adhere well to object boundaries (as shown in Figure 2.2).

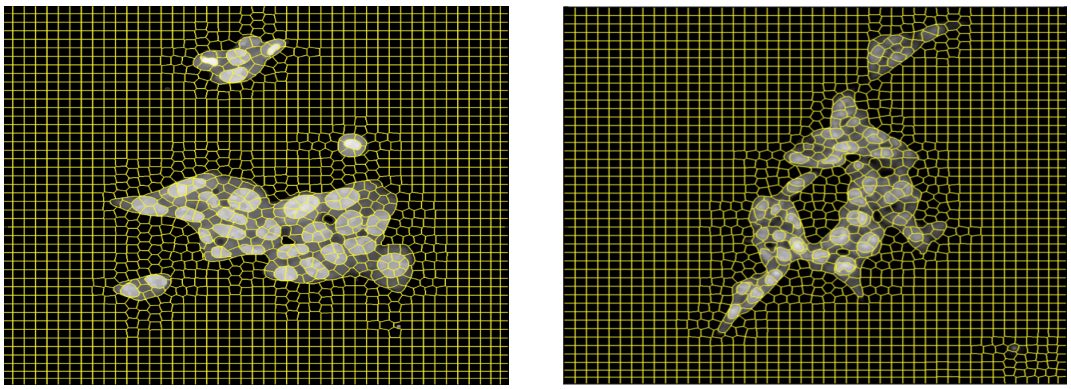


Figure 2.2: Sample result of the SLIC superpixel method. SLIC algorithm is widely used to generate locally consistent patches effectively. It is clear that the oversegmented regions adhere well to the boundary of objects.

Region merging methods based on a graph representation have also been proposed for over-segmented segments ([28–31]). A notable example is the normalized cut [29], where the original region merging problem is transformed into a graph segmentation problem. In their framework, the global criteria were proposed to segment the graph based on total dissimilarity between the different groups as well as the total similarity within the groups. He et al. [31] extended the algorithm by incorporating spatial and concave contour information, which enabled the detection and segmentation of a large number of cells in 2D images. Similarly, Al-Kofahi [32] adapted a graphical cutting-based method to perform segmentation of dense nuclei in cellular images. While these methods can achieve decent segmentation on specific datasets, it is difficult to generalise these methods to a large number of cellular images. More importantly, it is difficult to incorporate any prior knowledge.

Edge-based method. While region-based methods based on homogeneity criteria aim to detect regions with similar properties, edge-based techniques focus primarily on the delineation of object contours, which attempt to find discontinuities in the image. Two major tasks for edge-based methods are edge detection and edge linking. One of the most popular edge-based techniques is the Canny edge detector [33]. They first calculated and stored gradients in four different directions, then set a high threshold for the gradient to identify pixels that definitely belong to the object contour, and a low threshold to find pixels that might belong to the object contour. Finally, at the edge linking stage, an edge map was generated based on the pre-defined linking rules. In terms of cell segmentation, these edge detection methods also have some drawbacks. One of the problems is that this method may lead to broken edges and some texture edges inside the nucleus [34]. On the other hand, these methods are always based on the assumption that pixel properties, such as intensity, colour, and texture, can change abruptly from region to region. However, it is difficult to detect significant changes in regions with touching or overlapping cells. As a result, these methods can only achieve promising performance on simple cell imaging without overlap and complex textures.

In summary, the traditional data-based approaches discussed above, including thresholding, region-based and edge-based techniques, all have significant drawbacks in cellular images. Thresholding is simple, fast and cost-effective to implement, while the resulting segmentation can be scattered and incoherent. Region-based approaches address this problem, but the localisation of regions depends heavily on the selection of seed points, the setting of growth criteria, and the order in which pixels are processed, which is prone to over-segmentation. Edge-based approaches also have the problem of requiring post-processing to connect disjoint edges and eliminate undesirable textured edges. Finally, a common disadvantage of the three different types of data-based segmentation methods is the lack of advanced object modelling and the inability to incorporate a prior knowledge for the global control of the segmentation process.

2.1.2 Model-driven Methods

Rather than relying on data from an image, some segmentation methods, known as model-driven methods, leverage prior knowledge, including shape, size, or other attributes, to segment target objects. Template matching and deformable contours are two widely explored modelling techniques.

Template matching. The basic idea of template matching is to define a template of the target object in the dataset. For each input image, the template is shifted over the entire image, while calculating the similarity between the template and the underlying part of the input image. Finally, the regions with high correlation values can be detected as the segmentation result for the target object. The Hough transform [35] is proposed to detect predefined shapes such as lines, circles, ellipses, etc. In cell segmentation, the Hough transform is also effective in detecting the position and orientation of the nearly circular nuclei in a cell image [36]. In addition, the Hough transform can also be combined with a watershed algorithm to localise seed points in the nuclei of cells. [37] [38]. However, most traditional template matching techniques are not robust to the effects of rotation, scale, shape, appearance or illumination. To solve these problems, Liu et al. [39] propose an algorithm for multi-scale analysis based on filter design. They first constructed a filter library consisting of circles of different sizes to locate the nuclei in the image, and then refined the segmentation results at different smoothing levels. They also demonstrated that the proposed method achieved the best nuclear segmentation accuracy compared to the four state-of-the-art techniques.

It should be noted that pre-defined templates for static shapes are inflexible and unable to adapt to different target objects in an image. Therefore, Yuille et al. [40] proposed deformable templates that can be used to locate target objects with complex shapes. In these algorithms based on deformable templates, extra space is usually left to repeatedly adjust the pre-defined template to fit the complex contours of the object in the image. For example, a few parameters are used to describe the general shape of the object, then iteratively search for the best-fitting parameter

values based on the appearance of the object. These moving contours, also known as snakes, can be roughly divided into three different representations: point, geodesic, and parametric snakes. The details are explained in the next few paragraphs.

Point-snakes. Point-snakes [41] is based on the most basic representation of discrete curves, which is represented by a collection of adjacent labelled points and evolving along the direction of a negative gradient of image energy. From the mechanics of these algorithms, there are two main drawbacks that need to be mentioned. First, the representation contains a large number of parameters, leading to high computational costs and poor generalisation; second, the points defining the shape are usually simply connected by straight lines, making the method unsuitable for curve boundaries such as circles, unless numerous model points are selected.

Geodesic snake. Geodesic deformable model was proposed based on curve evolution and level set theory. The curve can evolve with the image data to detect the object boundary. Since the evolution is independent of parameters, the evolving curve can be represented as a level set of a high-dimensional function, i.e., a level set function, and thus topological changes are handled naturally. Currently, this deformable model has been widely used in the computer vision and machine learning community [42–44] since it does not require complicated parameters and does not suffer from topological constraints. Compared to traditional snake models, this approach is highly robust to noise and enables the segmentation of complex objects, such as sharp corners or cells with blurred edges. More importantly, this method only relies on the geometric measures (e.g., the unit normal or curvature), thus it does not require any labelled landmark in the image. The approach has been successfully extended to cell segmentation and tracking in 2D and 3D images [45, 46]. However, this method may make shapes difficult to constrain and lead to overfitting of the real biomedical images. In addition, the optimisation process is computationally expensive due to its high-dimensional representation method [47].

Parametric snakes. The third class is the parametric snake proposed by Staib and Duncan [48], where the object contour is defined by a set of parameters to ensure smoothness and continuity of the contour. Compared to point snakes or geodesic snakes, it is easier to incorporate prior knowledge of shapes and requires less computational complexity in the optimisation process. Hence, it is particularly suitable for the segmentation of objects with known or reproducible shapes that can be described naturally by parametric representations. Zimmer [49] presented a parametric deformable model to segment and track cell dynamics from in vitro video microscopy data. Based on parametric active contours, they employed a gradient vector flow (GVF) model to address important difficulties in cell imaging, such as touching cells, low-contrast boundaries, and inhomogeneous illumination. The proposed technique greatly enhanced the robustness of cell segmentation, with some improvements in locating blurred cell boundaries. However, their method relied strongly on manual initialisation, i.e., it only tracked cells that were manually initialised in the first frame and were unable to detect cells that entered the scene later in the sequence. While the parameters can be adjusted to fit precise object boundaries in most reported methods, similar to Zimmer's work, initialisation was still critically important as the adjustment of these parameters was limited.

2.1.3 Deep learning-based Methods

Deep learning methods have attracted a great deal of attention due to their superior performance in many computer vision tasks, including image classification [50, 51], semantic segmentation [52, 53] and object detection [1, 54]. In contrast to traditional image analysis pipelines that elaborately design hand-crafted features, advanced deep learning techniques now allow us to directly extract a representative set of features from the data that are highly relevant to the task at hand. Recently, many efforts have been made to target cell segmentation in biomedical images, and Fully Convolutional Networks (FCNs) [52] in particular have shown great potential to address the need for quantification of live cell imaging data. As shown in Figure 2.3, FCN optimised standard CNNs by replacing fully-connected layers with convolutions

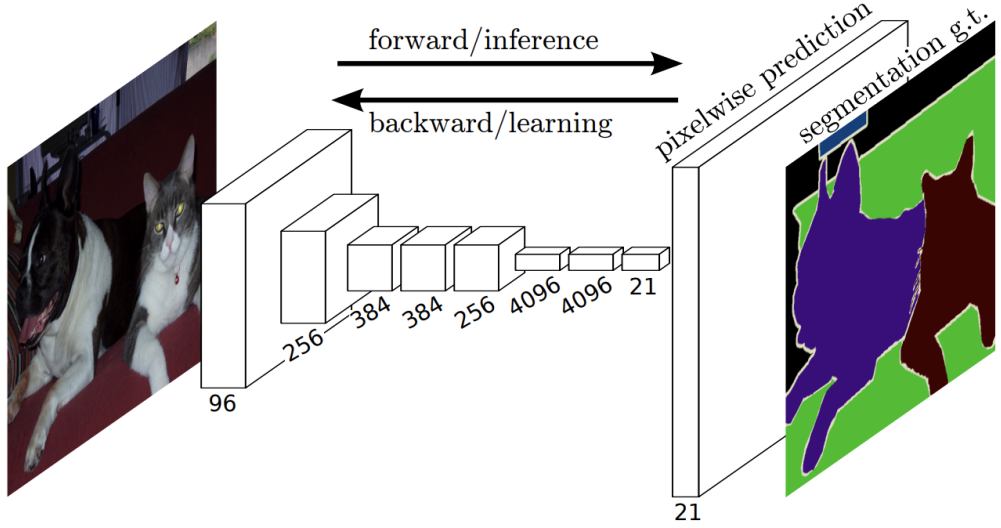


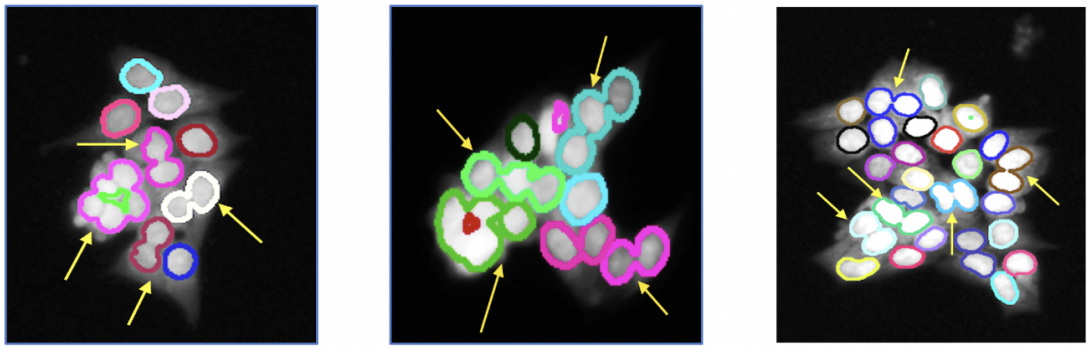
Figure 2.3: The overall structure of a fully convolutional network (FCN) [52].

of 1×1 kernel, and restoring the spatial resolution with learnable deconvolution layers. Therefore, the proposed FCNs can take an input of arbitrary size and generate an output of the corresponding size with end-to-end, pixel-to-pixel training. Xie et al. [1] addressed microscopic cell counting and detection problems with FCNs and demonstrated that networks trained only on synthetic data can be generalised to real microscopic images. Zhao et al. [55] proposed a pyramid-based FCN to segment cells with low-contrast and irregular shapes. In their architecture, high-level FCNs generate coarse cell segmentation masks to cope with low contrast between intracellular regions and backgrounds; low-level FCNs generate segmentation masks that pay more attention to cell details to tackle irregular cell shapes.

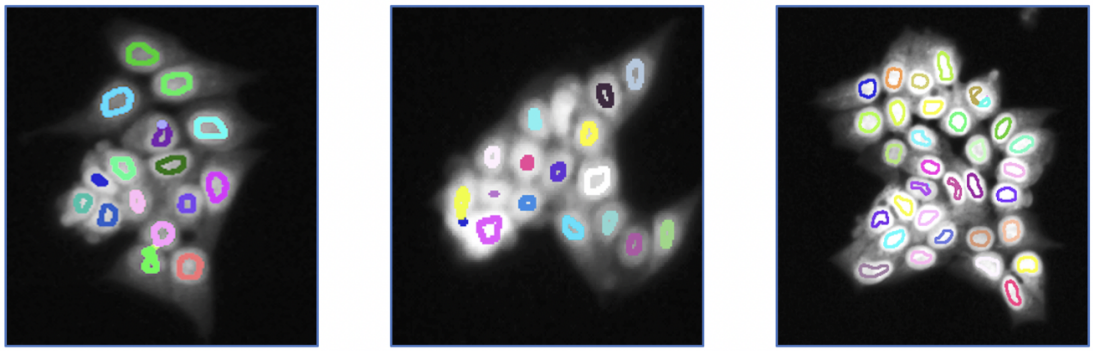
Inspired by fully-convolutional networks, Ronneberger et al. [5] proposed a deep convolutional neural network consisting of convolutional layers and deconvolutional layers with skip connections to combine low-level and high-level information. To address the challenge of touching cells, they proposed a weighted loss function in which a large weight was assigned to edge pixels between touching cells. As a result, the network won the ISBI Cell Tracking Challenge 2015 by a large margin of 46% over the second best algorithm. U-Net, an evolutionary variant of FCN, has been successfully employed for numerous segmentation tasks in biomedical image analysis.

Wang et al. [56] proposed a two-stage deep framework for high-resolution 3D data, which concatenated two modified U-Nets. Böhm et al. [57] described an extension of U-Net that elevated the label space from 2D to 3D and created a non-overlapping representation for each instance in 3D. However, this approach is obviously sensitive to the designed label transformations and requires extensive efforts to fine-tune its parameters. Raza et al. [58] extended U-Net by adding additional convolutional layers in the downsampling path to bypass the max pooling operation in order to learn features that were ignored in the maximum pooling operation, making it robust to noisy data. In addition, Zhou et al. [59] found two limitations of the U-Net architecture: (1) the optimal depth of the encoder-decoder network may vary from task to task; and (2) their skip connections are forced to aggregate only on the same scale feature map of the encoder and decoder subnetworks. To address the limitations mentioned above, they introduced UNet++, which consists of multiple U-Nets with different depths whose decoders are densely connected at the same resolution through redesigned skip connections.

Another limitation of these encoder-decoder architectures (FCN, U-Net) is that up-sampling in the decoder sub-network may not compensate for significant downsampling in the encoder sub-network, and thus detailed spatial information may be lost. Consequently, it is usually difficult to segment neighbouring cellular structures on biomedical images. To address this challenge, many segmentation techniques applied adaptive weighting schemes in the loss function to enhance the nuclear boundary pixels [5, 60]. Unlike previous binary classifiers that only discriminate nuclei against the background, the segmentation models with a weighted loss function can predict the inner region of the nuclear, the nuclear contour and the background at the same time. The example experimental results of U-Net trained using an adaptive weighting scheme are shown in Figure 2.4. Therefore, although this adaptive weighting scheme improves and stabilises the inner nuclei and boundary prediction, it is highly sensitive to the pre-defined weights in the weighted loss function.



(a) The weight of edge pixels was assigned to 1.



(b) The weight of edge pixels was assigned to 2.

Figure 2.4: The example experimental results of U-Net trained using an adaptive weighting scheme. When the weight of edge pixels was assigned to 1, the network identified a cluster of cells as an object. When the boundary weight was 2, most of the pixels inside and the cytoplasm near the boundaries were also predicted as boundaries.

Other contour-aware methods exploit information about the contours or perceptual boundaries to separate the attached objects. To name a few, Chen et al. [61] proposed a multi-task learning framework for instance segmentation that inferred the segmentation results and the complementary information of boundaries simultaneously. Oda et al. [62] designed a Boundary-Enhanced Segmentation Network (BESNet) for cell segmentation on histopathological images. Unlike the standard U-net, BESNet utilised two decoding paths, where an additional boundary decoding path can be used to enhance the performance of the entire cell segmentation, trained using the boundary labels of the annotated cells. To incorporate the boundary information, the feature maps in the boundary decoding path and the main decoding path were directly concatenated for the final cell segmentation. A

recent method, the Contour-aware Informative Aggregation Network (CIA-Net) [63], extended this idea by proposing a multi-level information aggregation module that allowed the two task-specific decoders to collaboratively refine the details of the nuclei and contours. Khoshdeli and Parvin [64] performed boundary-based and region-based segmentation using two independent encoder-decoder networks. The results of these two networks were then fed as input into another network for nucleus segmentation. In general, while these deep networks can produce promising segmentation results, a large scale of training data with corresponding ground truth label maps is required as the networks typically have millions of parameters to be optimised. In the biomedical domain, however, it is a major challenge to obtain such a large training set with manually-labelled ground truth maps due to the following three factors: (1) manual annotation requires a high level of expertise; (2) pixel-level annotation of large biomedical images is time-consuming and laborious; and (3) there is considerable intra- and inter-observer variability.

As a result, there has been significant interest in semi-supervised and self-supervised learning techniques, which aim to exploit a large amount of readily available unlabelled data to improve performance while reducing the manual annotation effort. For example, Bai et al. [65] proposed a semi-supervised learning approach for cardiac MRI image segmentation, in which not only labelled images but also unlabelled images were utilised in the training of segmentation networks. Gu et al. [66] constructed forest-oriented superpixels to augment the standard random forest classifier to take advantage of huge unlabelled data in biomedical image segmentation. Self-supervised techniques have also been investigated in the field of biomedical image segmentation but to a lesser extent. The underlying idea is to assign artificial labels to unlabelled biomedical images through a pretext task, and then use the resulting supervision to pre-train a deep network for the target task that lacks sufficient labelled training data. Tajbakhsh et al. [67] studied the impact of three pretext tasks (rotation, reconstruction, and colourisation) in four different medical imaging applications for both 2D and 3D biomedical images. Bai et al. [68] leveraged the rich information encoded in the cardiac MR images

and proposed a novel pretext task based on anatomical positions. Self-supervised feature learning provides a viable solution for limited annotated data in biomedical imaging; however, the learned feature representations are not as effective as those learned through strong supervision. Moreover, while information from unlabelled data can be fully involved in these methods, some low-quality unlabelled data may be unsuitable for feature learning, and it is still an open question on how to select appropriate unlabelled data.

2.2 Biomedical Image Classification

In this section, we provide a brief overview of the development of deep learning strategies for biomedical image classification. We first present the technical background of applying general deep learning architectures to image classification problems in computer vision. We then focus on the use of single-level neural networks and two-level deep learning approaches to tackle the classification problem of high-resolution biomedical images.

2.2.1 Image Classification in Computer Vision

In the past few years, deep CNNs have made tremendous progress in the field of large-scale image classification [50]. Krizhevsky et al.[50] propose a new CNN based architecture, named AlexNet, which was the first application of deep learning networks to large-scale image classification. In 2012 ImageNet Classification Challenge, AlexNet achieved record-breaking performance with a top-5 error rate of 16.4%, which is about 10% lower than the second team using the traditional algorithm. As shown in Figure 2.5(a), AlexNet is an 8-layer convolution neural network consisting of five convolutional layers and three fully-connected layers, with the last layer being classified by a 1000-way softmax classifier. Central to their success, they used Rectified linear units (ReLU) rather than the traditional sigmoid and hyperbolic tangent activation functions, and employed Dropout [69] as a regularisation method to overcome overfitting problems. Due to the AlexNet model, the field of computer vision has been revolutionised and, as a result,

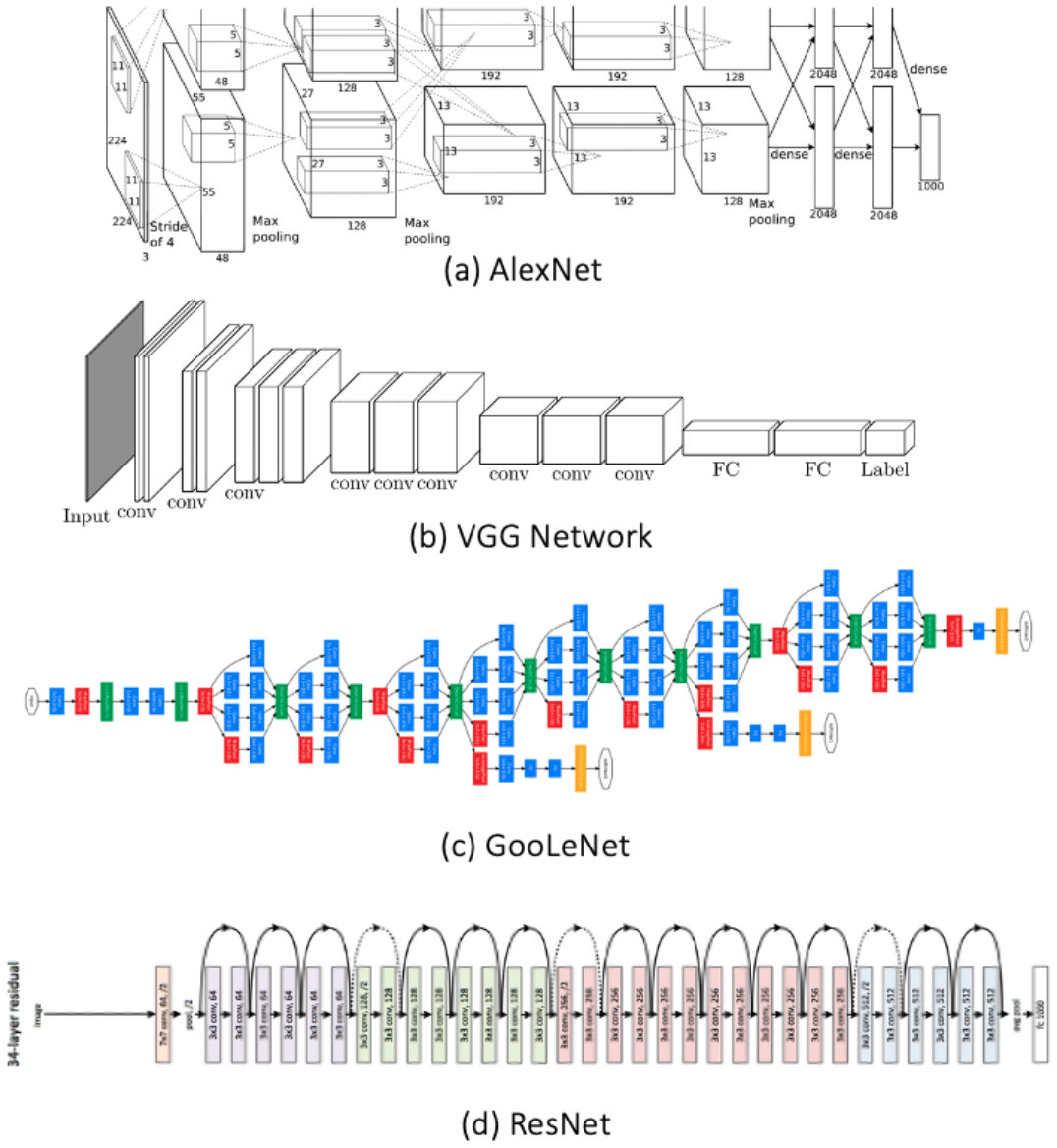


Figure 2.5: Architecture of AlexNet[50], VGG Network[51], GooLeNet[70] and ResNet[71]. The figures are taken from the original papers.

deep convolutional neural networks have become the dominant method for image classification-related applications.

Inspired by AlexNet, many recent works on image classification have attempted to improve the classification performance by modifying the architecture to deeper and wider networks [51]. In particular, Simonyan and Zisserman[51] proposed a deep convolutional network with 19 convolutional layers, known as VGGNets (Figure 2.5(b)). Instead of using the large filters in AlexNet (11 and 5 in the first and

second convolution layers, respectively), they achieved the improvement by using multiple smaller-sized convolutional (3×3) filters throughout the network. Similar to the idea of VGGNet, Szegedy et al.[70] also introduced a very deep convolutional network, GoogleNet (Figure 2.5(c)), consisting of 22 layers. The enhancement of GoogleNet is primarily based on a multi-scale processing approach. In order to find the optimal local sparse structure of the image, they proposed an architecture called Inception module which consists of dimension reduction 1×1 convolution filters, 3×3 and 5×5 convolutional layers, and 3×3 max pooling layers. In this way, they achieved effective dimensionality reduction prior to the more computationally costly 3×3 and 5×5 convolutions, which were able to increase the width and depth of their network with the same computational cost.

Deep CNN architectures are generally more accurate and therefore produce better classification results than the previous shallow networks; however, researchers found that convolutional networks would encounter issues with vanishing or exploding gradients as the number of layers deepens, which significantly affects the performance of the deep networks. To address this problem, He et al.[71] introduced a new deep residual learning framework (ResNet). Instead of learning the mapping directly, they reformulated the architecture of the network, forcing it to learn the residual function with inputs from the previous layer. To achieve the residual learning, shortcut connections are incorporated into layer-to-layer connections, which perform identify mappings whose outputs are added to the output of the stacked convolutional layers (Figure 2.5(d)). This allowed gradients to be back-propagated more easily, and thus classification accuracy can be improved as the network goes deeper.

2.2.2 One-stage Deep Classification of Biomedical Images

In the last few decades, image classification has been intensively studied, in which a large number of deep learning-based methods have been proposed in the literature. Meanwhile, the recent advances in the field of medical imaging have enabled us to acquire a large number of high-resolution images from a wide range of clinical imaging modalities, such as microscopy, ultrasound, computed tomography (CT),

Magnetic Resonance Imaging (MRI) or digital pathology. As a result, with the large amount of relevant information contained in today's biomedical images, deep learning techniques have been successfully used as a tool to enable efficient workflows in clinical practice. Among deep learning techniques, deep convolutional networks are most actively used in a variety of biomedical image analysis tasks. It covers segmentation, retrieval, abnormality detection, disease classification and computer-aided diagnosis. In this section, we specifically summarise the main work on applying single deep convolutional networks to biomedical images.

Recent advances in deep convolutional networks have dramatically improved the performance on various biomedical image classification problems, such as breast histology image classification ([72–75]), skin cancer identification ([76–78]), lung nodule classification ([79–81]) and many other similar applications. To apply deep convolutional neural networks (CNNs) to biomedical image classification, the most straightforward solution is transfer learning, where powerful CNN models pre-trained on natural images are fine-tuned on biomedical images. Rakhlin et al. [82] presented a histological microscopic image analysis approach for the classification of breast cancer types. They utilised several pre-trained deep convolutional neural network (CNN) architectures (VGG16, ResNet-50 and Inception-V3) for effective feature extraction and gradient boosted trees for final classification. Ferreira et al. [83] downscaled the high-resolution images to 224 x 224 pixels and fine-tuned an Inception ResNet V2 for classifying breast histology images into four classes: normal tissue, benign lesion, in situ carcinoma and invasive carcinoma. In [84], Brancati et al. proposed a fine-tune strategy of ResNet [71] for the multi-classification of breast cancer histology images and obtained prediction results by combining three ResNets with different numbers of layers. In [2], Kone et al. presented a hierarchical structure of the ResNet [71] model, which progressively classifies the images from carcinoma and non-cancer, and then into the four groups described above. These results demonstrated the potential application of deep convolutional networks in the classification of histological microscopic images.

In addition, the work of Esteva et al. [85] set the cornerstone for applying deep learning techniques to skin lesion diagnosis, as their deep networks enable dermatologist-level classification performance. In their design, a pre-trained deep convolutional network (Inception V3) was fine-tuned on 129,450 clinical dermoscopy images and its predictions were compared to the performance of 21 trained dermatologists. Matsunaga et al. [86] improved the melanoma classification by aggregating multiple pre-trained CNNs into a single framework where the finial predictions were calculated based on the weighted sum of the individual networks. Similar ensemble learning strategies were also employed in [87], [88] and [89] to boost the classification of melanoma, nevus and seborrheic keratosis. Yang et al. [76] proposed an end-to-end melanoma recognition framework that used the segmented lesion masks to guide the classification performance through a new pooling operation named region average pooling. The proposed framework consists of a regular ResNet-50, followed by two branches: segmentation branch and classification branch. With an explicit combination of segmentation and classification, they achieved state-of-the-art performance on ISBI 2017 test data [90] without using additional images. Also, the classification performance of skin lesion can be improved by incorporating as much as possible training data [91, 92].

Similarly, pre-trained CNNs have also been applied to different computed tomography (CT) images. Shin et al. [3] explored and examined different CNN architectures on two specific computer-aided diagnosis (CAD) problems: lymph node detection and interstitial lung disease classification. Gao et al. [93] fine-tuned all layers of a pre-trained AlexNet to classify interstitial lung diseases in holistic CT images. They demonstrated the state-of-the-art classification performance on a publicly available dataset and showed strong transfer learning ability of the DCNN. In [79], Shen et al. presented a multi-crop CNN for the problem of lung nodule malignancy suspiciousness (the likelihood of nodule malignancy) using CT images. In this end-to-end architecture, they proposed a new multi-crop pooling operation that can crop multiple different regions from the previous convolutional feature maps. In [80], Xie et al. proposed a novel algorithm for lung nodule classification on chest

CT images, which jointly used texture, shape and deep convolutional features at the final decision level. Motivated by class activation mapping, Lei et al. [94] developed a soft activation mapping module where the importance of each feature map in the last convolutional layer is obtained via average pooling followed by a fully connected layer. This design enabled the network to refine the fine-grained lung nodule features, achieving state-of-the-art classification accuracy on LIDC-IDRI [95] dataset.

Generally, this is a remarkable achievement of one-time fine-tuning—simply fine-tuning a pre-trained CNN model with available training samples. These deep learning methods have the ability to learn multiple levels of feature representation from original raw images, as there is no need to design hand-crafted features. With respect to architecture, VGG networks, ResNets or Inception networks pre-trained on ImageNet play a dominant role in biomedical image classification based on transfer learning strategies. The VGG network follows the layout of a traditional CNN model: a sequence of convolutional, maximum pooling, and activation layers prior to a number of fully connected classification layers. The ResNet a stack of several residual learning blocks equipped with multiple skip-connections between the identity mapping of the input and the output. Finally, the Inception network features convolution kernels of different sizes as well as multiple max pooling layer. However, some challenges still remain with these one-stage deep learning approaches. Firstly, these methods only extract global features from down-scaled low-resolution images, which may result in sub-optimal results in biomedical applications. Second, high performance is primarily achieved by fine-tuning a standard CNN model and increasing the size of the dataset, rather than by specific model design. Finally, relatively small datasets and expensive expert annotations still pose significant challenges to the automatic classification of biomedical images.

2.2.3 Two-stage High-resolution Classification of Biomedical Images

Imaging data of biomedical samples are typically characterized by large size and high resolution. Moreover, training a CNN on high-resolution biomedical images

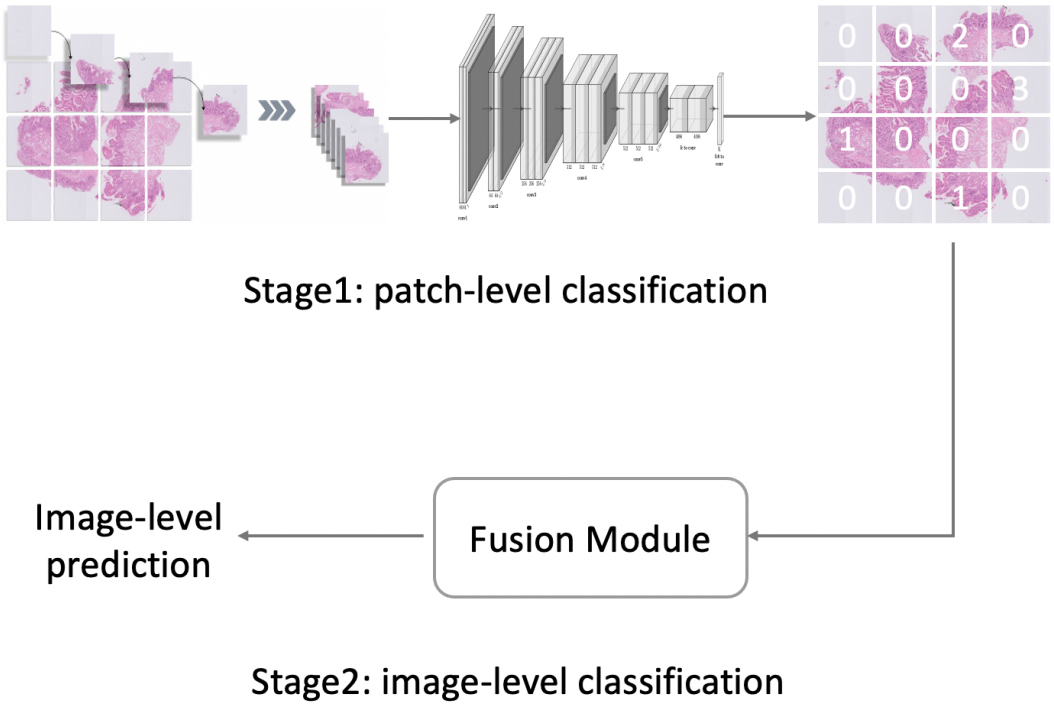


Figure 2.6: The overall structure for two-stage classification of biomedical images. CNN cannot be applied to very high-resolution images due to memory limitations and high computational costs. Therefore, a two-stage learning strategy is widely used for high-resolution biomedical image classification. In the first stage, the high-resolution images are partitioned into locally equal-sized patches, which can produce an extremely large number of small patches to train the patch-level CNNs. In the second stage, the patch-level classification results are aggregated into image-level predictions.

requires high computational costs and a very large amount of memory, which is intractable in most cases. Most approaches tackle this problem by reducing the space size of the image which allows the downsampled version to be stored in memory and fed directly to the deep network. However, downsampled images run the risk of losing discriminatory features in the high-resolution version such as nuclei or lymphocyte information. Alternatively, several two-stage methods have been proposed in the literature to preserve the local high-resolution features [73, 74, 96]. The core ideas of these methods are much the same. As shown in Figure 2.6, the high-resolution image is pre-processed and divided into local equal-sized patches, which can generate a very high volume of smaller patches to train the deep network. The extracted patches are classified by a patch-wise network, and the

final image-level classification is obtained by fusing patch-level information.

In the literature, a number of studies have investigated fusion methods for aggregating patch-level labels to image-level labels. Vang et al. [73] proposed a deep learning framework for multi-class breast cancer histology image classification. They used pre-trained Inception (GoogLeNet) V3 model to train a patch-level classifier, and patch-level results were then passed through an ensemble fusion framework containing majority voting, gradient boosters (GBM), and logistic regression to obtain image-level predictions. To incorporate the contextual information, Nazeri et al. [96] also proposed a two-stage pipeline for breast cancer microscopy image classification based on two consecutive convolutional neural networks. The first network was trained on image patches to extract the most salient patch-level features while the second image-wise CNN learned global information for the classification of the high-resolution image. Shaban et al. [97] extended this idea by exploring several state-of-the-art architectures for the second image-wise CNN. In this way, the feature learning and aggregation are explicitly stacked together for the learning of large spatial context. However, the pre-extracted patches may ignore the long-range correspondences and suppress the important fine-grained information, such as the density or shape of objects. To this end, Yan et al. [75] proposed a different fusion module for integrating the predictions of small patches. Instead of a CNN, they utilised a recurrent neural network (RNN) to fuse the contextual information of patch-level convolutional features where short-term and long-term spatial correlations between patches can be both considered.

In addition to fusion methods, how to extract the patch-level information in a meaningful way still needs to be investigated carefully. Gessert et al. [98] introduced a patch-based attention method to identify the informative patches and aggregated the patch-level features based on the learned attention weights. Golatkar et al. [74] optimized the patch-level classification by nuclei-based patch selection where only patches with high nuclear density were used for patch-wise network training. Similarly, Zhang et al. [78] presented a multi-scale strategy for patch extraction on dermatoscopy images, which extracted patches of different scales from only the

central part of each image. Diaz [99] designed an extra segmentation network to incorporate the segmentation information into the classification pipeline of skin lesions. This segmentation network generated binary masks outlining the lesion areas in the dermatoscopy images. Although this method allowed for the extraction of a set of patches surrounding the lesion area, it might cut off the relevant part of the classification task owing to segmentation errors.

Multiple instances learning (MIL) for biomedical image classification. All of the above methods are based on the assumption that all individual patches that are cropped from the same image have a consistent label with the original high-resolution image. However, patch-level labels do not always match image-level labels in the context of biomedical imaging. For example, when we classify cancer cases into grades and subtypes, patches from the same high-grade image may have a mixture of high- and low-grade characteristics. To determine the image grade, we need to find discriminative patches that show typical visual features of high-grade cancers and ignore non-discriminative patches that contain only visual features of low-grade cancers. In this case, a large body of work on the problem of classification of biomedical images has introduced multi-instance learning (MIL), which enables patch-level classification in the presence of only image-level labels.

MIL was first introduced by Dietterich et al. [100] for drug activity prediction, and has achieved great success in many biomedical imaging applications over the past 20 years [101, 102]. It is one of the latest variants of the weakly supervised paradigm, in which training data (instances) are arranged in sets, often called bags. Unlike supervised learning, which labels each single instance in a dataset, the MIL algorithm is designed to use only global bag-level labels for both bag- and instance-level classification tasks. Specifically, each bag $X_i = \{x_{i1}, ..., x_{iM}\}$ is associated with a target label $y_i \in \mathcal{Y} = \{-1, 1\}$, while individual labels of the instances $y_{ij} \in \mathcal{Y}$ for $j = 1, ..., M$ are not provided. The standard MIL assumption for binary classification states that a bag will be labelled positive if and only if it contains at least one positive instance. A bag is labelled negative if and only

if all instances are negative. Let $h(x_{ij}) : \mathcal{X} \rightarrow \mathcal{Y}$ be a instance-level classifier that returns labels for each instance. The bag-level classifier $H(X_i) : \mathcal{X}^m \rightarrow \mathcal{Y}$ can be given by: $H(X_i) = \max_j h(x_{ij})$.

MIL seems well-suited for biomedical imaging, in which it is not computationally feasible to process high-resolution images consisting of billions of pixels. Moreover, in biomedical fields, it would be prohibitive and tedious to obtain pixel-level annotations, which poses significant difficulties in building large-scale datasets. Therefore, it is preferable to divide a large biomedical image into smaller patches that can be further thought of as a bag with a single label. In particular, the MIL-based CNNs have attracted a great deal of interest in the classification of biomedical images, where each image typically corresponds to a bag, and each cropped patch is referred to as an instance. Hou et al. [103] presented an EM-based MIL algorithm to iteratively identify discriminative patches in high-resolution images for a two-level patch-based CNN applied to subtype classification of lung cancer cases. However, it is not feasible to apply EM-based methods to whole-slide images in clinical practice, due to the small proportion of abnormal instances and the significant computational cost. Yan et al. [104] designed a multi-stage deep framework for biomedical image classification within a MIL setting and applied it to the bodypart recognition problem. In the first stage, a multi-instance learning strategy based on max pooling was utilised to discover the discriminative local information of different body parts while selected local patches were used in the second stage to boost the learning of pre-trained CNNs. Kraus et al. [102] developed a deep convolutional neural network equipped with MIL for the classification and segmentation of microscopy images with only image-level annotations. For facilitating MIL aggregation over a large number of instances probabilities, they proposed a new MIL pooling function, called Noisy-AND pooling, which assumes that a bag is positive if and only if it contains positive instances above a certain threshold. In a latter study, Zhu et al. [101] proposed an end-to-end deep multi-instance network that brings MIL assumption into the problem of unannotated whole mammogram classification. In this scenario, each mammogram is treated as a bag, and all extracted patches can be treated

as instances. Instead of using max pooling which only considered the patch with maximum probability, they converted the MIL assumption into a label assignment problem and designed three MIL loss layers: max pooling, label assignment and sparsity loss to explore information from other patches.

2.3 Relevant Advances in Deep Learning

2.3.1 Attention Mechanisms

Recent work has shown that although the CNNs have only been trained on image-level labels, they have significant capabilities in naturally directing attention to salient regions of the input image [105, 106]. A number of methods [105, 107] have been proposed to visualise CNN predictions based on the concept of saliency in images. Specifically, the class saliency map was proposed by Simonyan et al. [107] to highlight the important regions in a given image that have an active impact on the classification result. These image-specific class saliency maps are obtained by calculating the gradient of the class score, with the aim of querying the already-trained classification network and visualising the salient regions associated with a particular class. However, saliency maps are not used directly in training networks and generally do not serve to improve performance or reduce computation. To achieve this, attention maps [108–110] extend saliency maps by enforcing the classification of important regions during the network training process. The attention models can be incorporated into multiple components of deep networks and improve the performance by increasing the weights of specific parts of the input data.

A large body of research has demonstrated the promise of attention mechanisms in addressing discriminative fine-grained feature learning. Motivated by these recent studies, this thesis investigates attention-based fine-grained classification methods for biomedical images in Chapters 4 and 5. This section aims to review the existing attention-based techniques in computer vision and biomedical image classification.

Attention mechanisms in computer vision. Attention mechanism has been extensively studied in the fields of computer vision and natural language processing, such as image inpainting [108], image and video captioning [111], semantic segmentation [110] and machine translation [109]. Recently, various methods incorporate attention mechanism to enhance the feature representation ability of CNNs, which demonstrated that the attention mechanism is effective in highlighting the salient regions and discarding the irrelevant information in computer vision tasks.

A desirable feature of the attention mechanism applied to CNNs is flexibility and architectural independence, i.e., the attention module can be easily adapted to any CNN architecture. Based on these properties, Rodriguez et al. [112] proposed a generic architecture-independent attention module that can be applied at different depths to highlight the discriminative features of the classification. Each attention module consisted of three main submodules: (i) the attention heads, which define the most relevant regions, (ii) the output heads, which generate class probability scores based on the attended information, and (iii) the confidence gates, which output a confidence score for each attention head. In this way, the model learns to select the most meaningful attention heads in a given attention module. Similarly, Wang et al. [113] proposed a residual attention network that is built by stacking multiple attention modules in which a bottom-up top-down structure is used to generate attention-aware features that highlight the most discriminative regions of the convolutional feature map. In each attention module, the former bottom-up path gathers information from the feature map, while the latter top-down path combines attention weights with the original feature map. Similar to the idea in the original ResNet, they added a residual learning mechanism to the naively stacked attention modules, which allows the network to maintain good performance as the layer goes deeper. While such a modular design is fast, simple, and independent of the network architecture, the computational cost of these modular attention mechanisms is significantly higher due to the large number of additional learnable layers.

Many studies have concentrated on the objective of reducing complexity and computational cost of an attention module. Recent approaches to attention-based

learning can be divided into two categories: the first is global information modelling using a fully-connected layer. The second types use self-attention mechanisms to model pair-wise relationships. Regarding the global information modelling, Hu et al. [114] introduced a simple and lightweight module, Squeeze-and-Excitation (SE) block, to investigate the relationship between channels. The proposed mechanism allows the network to adaptively recalibrate feature responses by explicitly modelling interdependencies between channels. The first step applies global average pooling to squeeze the feature maps into one-dimensional channel descriptors. In the second step, a fully connected neural network with two hidden layers is used to automatically learn channel-wise nonlinear interactions. Woo et al. [115] extended this idea to exploit attention mechanisms along with both spatial and depth directions. They designed the Convolutional Block Attention Module (CBAM), which has two sequential sub-modules: a channel attention module based on the SE block and a spatial attention module that applies average pooling and max pooling operations along the depth direction. While the additional computational overhead can be greatly reduced in these SE-based methods, the input feature maps are first compressed into a 1D feature vector, resulting in a loss of position correspondence. To this end, the recently proposed Gather-Excite (GE) network [116] builds on the SE network to further explore mechanisms for incorporating context into the overall deep network architecture. In their design, the gather operator first aggregates contextual information across large neighbourhoods in each feature map. The resulting context information is then passed along with the original input tensor to an excitation operator that produces a new tensor of the same dimension as the original input.

Instead of learning attention weights by modelling global contexts, many approaches take advantage of the self-attention mechanism to model the pairs-wise dependencies. Wang et al. [117] presented a widely-used non-local network (NLNet) that computes the attention weight of the query position based on the pixel-level pairwise relationships between this query position and all possible positions in the feature maps. This allowed the network to capture the long-range

contextual information between distant pixels. Huang et al.[118] accelerated the non-local operation by incorporating recurrent criss-cross modules and a category consistent loss. Unlike NLNet, which measures the pairwise relationships for each query location, the proposed criss-cross attention module only collects sparse contextual information in horizontal and vertical directions. An additional category consistent loss was used to force the criss-cross attention module to generate more discriminative features. Following this line of research, Zheng et al. [119] extended this idea to explicitly model the inter-channel correspondences. They first obtained bilinear features that represent the spatial correspondence between channels and further integrated the spatial relationships into the feature map by performing dot production. Fu et al. [110] proposed a dual attention module that combines the above ideas to capture features dependencies in both spatial and channel dimensions. They built two parallel self-attention modules: one is a position attention module and the other is for channel attention, where the outputs were concatenated to further strengthen the feature representations. Cao et al. [120] found that NLNet actually produces roughly the same attention map for different query locations, which leads to a waste of time and computation cost. To address this problem, they integrated the Squeeze-and-Excitation module into the NLNet and designed a global context modelling framework. Specifically, it is a three-step attention module. (a) a context modelling module, which uses a global average pool to aggregate features together; (b) a bottleneck transformation module, which uses a 1x1 convolution, a ReLU, and a 1x1 convolution to capture channel-wise interdependencies; and (c) a fusion function, which merges global context features into the original features. However, most of these methods only captured the pairwise relationships among existing convolutional features, which is not effective enough to discard the irrelevant information on small training datasets.

Attention mechanisms in biomedical image classification. Inspired by the success in natural image analysis, attention mechanisms have also been exploited extensively in the context of biomedical image analysis, such as disease diagnosis

[121], organ detection [122] as well as tumour segmentation [123]. The attention mechanisms show a strong attention ability to highlight discriminative regions rather than irrelevant backgrounds, and there is a great deal of work on automatic classification and segmentation of biomedical images.

Since higher layers have better semantic abstractions than lower layers in deep convolutional networks, many classification and detection frameworks take advantage of the more abstract feature maps generated by higher layers to modulate the features of lower layers. Guan et al. [121] proposed a three-branch attention guided convolution neural network (AG-CNN) for thorax disease diagnosis on chest X-ray (CXR) images. First, they generated binary masks in the global branch to locate the discriminating region in the global image. Then, they cropped an informative region and trained it for patch-level classification in the local branch. Finally, the global and local information was fused in the fusion branch for final classification. However, only one prominent lesion or region of interest (RoI) was extracted and the method has limitations for applications where there may be multiple discrete RoIs. Schlemper et al. [124] introduced a grid attention mechanism consisting of multiple learnable 1x1 convolution layers and a grid resampler. They first extracted the global contextual information from the high-level feature maps and identified the salient regions at different scales based on the global context. The proposed attention module was evaluated on various biomedical image classification and segmentation tasks, such as fetal ultrasound image classification. While the attention mechanism is effective in improving the performance of CNNs on large datasets, the attention weights are learned through the use of extra learnable layers, which not only increases the computational cost, but also may result in overfitting on small training datasets. To address this problem, Zhang et al. [78] applied attention residual learning for skin lesion classification in dermoscopy images. The attention weights in their architecture are obtained directly from the classification network itself without extra overhead. Wang et al. [122] designed a feature guidance branch that compares low-level feature maps with a guidance map generated in the higher layer, thereby replacing the lower-score feature maps with the higher-score feature

maps. In addition to the unsupervised attention mechanisms, which are guided by the image-wise category labels only, Yang et al. [125] proposed to use the coarse annotation of the diagnostically relevant regions to guide the attention module for the classification of breast cancer microscopy images. In this way, the guided soft-attention mechanism guides the classification network to pay more attention to the diagnostically relevant regions. Similarly, Son et al. [126] collected the regional annotation of target findings in fundoscopic images, and trained a CNN to learn meaningful patterns for the classification task, guided by the area annotations.

Attention mechanisms in biomedical image segmentation. In addition to the automatic classification, researchers have already devoted a great deal of effort on incorporating attention mechanisms into the semantic segmentation of biomedical images. Target structures on medical images typically exhibit intra- and interclass diversity in shape, size, and texture. Conventional CNNs used for segmentation only have a local receptive field, which leads to potential differences between features corresponding to pixels with the same label due to the loss of long-range context information, especially when the network is trained on small datasets. Therefore, attention mechanisms are heavily used to capture long-range dependencies and further improve the discrimination of feature representations for pixel-level segmentation. Yi et al. [127] proposed an attention-based network for neural cell segmentation, which employs attention mechanisms in both detection and segmentation modules to enable accurate prediction of each cell's bounding box and its segmentation mask simultaneously. Specifically, the attention module forces the network to focus on information regions while suppressing irrelevant background, allowing the network to capture the tiny and slender structures of neural cells. Sinha and Dolz [128] introduced a multi-scale guided attention network for abdominal organ segmentation on magnetic resonance imaging (MRI). This multi-scale strategy employed a stack of dual attention modules [110] at different resolutions to filter confounding regions unrelated to the semantic representation of the target organ. In [129], a similar two-stage organ attention network was

proposed for the 3D volume of abdominal CT images where a simple attention module was adopted to filter redundant information and provide spatial attention. Shallow layers are responsible for finer spatial features and have more detailed information, while being weak in high-level semantic information. Deeper layers have larger receptive fields and capture global semantic information, but suffers from a loss of local details. In order to better leverage the strengths of each, many approaches adopt attention mechanisms to aggregate low-stage and high-stage information. Wang et al. [130] developed a novel 3D feature pyramid network equipped with attention modules to generate deep attentive features for automatic prostate segmentation in 3D transrectal ultrasound images. On each individual layer of the CNN, their layer-wise attention module selectively leverages complementary features integrated from all scales including low-level detail and high-level semantics to refine the feature representation of this layer, suppressing details not located in regions of semantic salience in the shallow layer and adding more boundary details in the deep layer. Rather than directly integrating high-level features with low-level details, Wang et al. [131] propose the distillation and edge attention modules that use only the information of shallow layers to refine the spatial information and restore image details of high-level layers. Also, a similar attention mechanism is adopted in a hierarchical aggregation framework [132] to adjust the convolutional responses for 3D left atrial segmentation. In addition to supervised learning, attention mechanisms are also introduced to bring unlabelled data into network training in semi-supervised learning strategies. Chen et al. [123] proposed an attention-based semi-supervised framework for medical image segmentation, in which they integrated an autoencoder with a U-Net-like network. This approach uses an attention mechanism to create labels for unlabelled training data and trains the autoencoder to reconstruct the segmentation based on these new labels. They incorporated the proposed learning scheme into two segmentation problems: brain tumour and white matter hyperdensity segmentation, demonstrating that the performance outperforms a fully supervised CNN baseline.

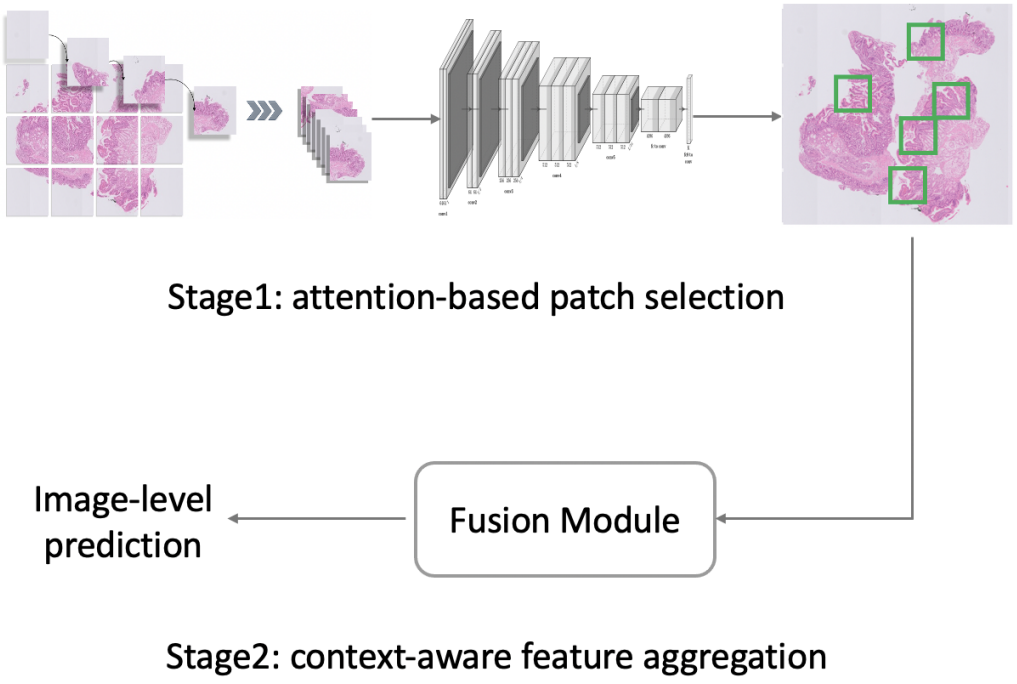


Figure 2.7: The overall structure for two-stage attention-based classification of biomedical images. In this paradigm, attention mechanisms are used to select informative patches in high-resolution images. In contrast to the standard two-stage framework, which produces patch-level classification results, this attention-based scheme highlights discriminative patches in the classification task.

Attention mechanisms in high-resolution image analysis. As discussed in Section 2.2.3, a two-stage learning strategy has been widely used to conduct high-resolution image analysis. However, it is challenging to apply these standard two-stage methods to whole-slide datasets, where discriminative patches only occupy a small partition of the image. More importantly, the large number of normal patch features may suppress discriminative information from informative patches, leading to inaccurate image-level label predictions. To tackle this problem, many methods utilised attention mechanisms to identify informative patches before aggregating the patch-level information, which enables the network to pay more attention to critical patches (Figure 2.7). As opposed to the aforementioned attention mechanisms, these patch-wise attention module tends to discover highly informative patches rather than discriminative locations within an image. For example, Gessert et al. [98] proposed a patch-wise attention mechanism for skin lesion classification,

which guides the network to focus on the lesion regions by incorporating the global and local context information. Similarly, Shaban et al. [97] designed a context-aware framework for grading the colorectal cancer histology images. They applied the attention mechanism to the high-dimensional patch-based representation and assigned high attention weights to important patches. To achieve high-resolution image classification using only image-level labels, some methods formulate the process of aggregating patch-level features into image-level predictions as a multiple-instance learning (MIL) problem, a weakly supervised framework in which each patch is considered as an instance and each whole slide image is treated as a bag, providing only bag-level labels. Ilse et al. [133] presented an attention-based MIL mechanism for weakly supervised histopathology image analysis, which allows the network to find the key instances. This trainable MIL pooling operator aggregates patch-level information based on the different weights of the instances within a bag, where a two-layer neural network is used to learn the attention weights. Lu at et. [134] extended this idea by incorporating contrastive predictive coding (CPC) to take advantage of the large number of unlabelled patches in the MIL framework. A latter study [135] was conducted to present an efficient two-stage framework for the automated gastric whole-slide image classification. The designed network first selected the discriminative patches based on the probability map of a pre-trained localisation network, and applied an attention mechanism to estimate the contribution of each selected patch for image-level label prediction. However, the process of discriminative instance selection requires pixel-level annotations to train the localisation network in a fully supervised manner. Recently, Rymarczyk et al. [136] enriched the mechanism of attention-based MIL with a self-attention mechanism, taking into accordant the pairs-wise spatial correspondence across instances. The study pointed out that the most important limitation of attention-based MIL is that this approach addresses all instances of the bag independently and ignores the correlations between the different instances. To address this restriction, they extended it by adding an additional self-attentive module before the attention-based MIL pooling operator to model the interconnected information at the instance level.



Figure 2.8: Fine-grained classification VS. general image classification. The blue box contains images of different species of birds, namely Scott Oriole and Orchard Oriole. The red box gives an images of a dog. Fine-grained recognition focuses on objects belonging to the same basic class, whereas general image classification distinguishes between basic-level categories.

2.3.2 Fine-grained Image Classification

In the past few years, deep convolutional networks have been shown to have superior performance when classifying basic image categories on large datasets (see Section 2.2.1). Unlike general image recognition, many real-life applications may involve the classification of subclasses belonging to the same basic class, such as species identification in biomedical image analysis. For example, in the blue box of Figure 2.8, there are two different birds, the Scott Oriole and the Orchard Oriole, which may be difficult for humans who are not experts to distinguish. However, general images are relatively straightforward to classify; for example, most people can easily identify the two in the blue box as birds and the one in the red box as a dog. Therefore, this task remains challenging for existing methods, as fine-grained image recognition should be able to locate and distinguish subtle visual distinctions in key parts within subordinate categories, especially when objects are not properly aligned, e.g., variations in background and pose.

Local annotations for fine-grained image classification. A large number of works address this problem by explicitly using manually annotated bounding boxes or even part annotations (e.g., bird heads, bodies) to depict part-based

representations of features. Early works take advantage of human annotations in both the training and testing phases. For example, Liu et al. [137] presented an example-based geometric model capable of detecting dog faces and locating their face parts from keypoint annotations for dog breed identification. Berg and Belhumeur [138] used part location labels to create a framework for building a library of highly discriminative intermediate features, namely "part-based one-vs-one features" (POOFs). Xie et al. [139] introduced a hierarchical partial matching (HPM) model for fine-grained bird species recognition, in which ground truth bounding boxes and part-level annotations are fully utilised to boost the classification performance. Branson et al. [140] showed an empirical study of pose normalisation schemes based on keypoint annotations for fine-grained visual categorisation. In a recent paper, Zhang et al. [141] proposed a top-down architecture consisting of two sub-networks: a part detection sub-network trained with part-level annotations and a recognition sub-network for image-level recognition.

Obviously, this setup of using part or keypoint annotations is demanding, and a more realistic setting would be to use only object bounding boxes to distinguish similarly fine-grained classes. Chai et al. [142] described a symbiotic system in which both foreground segmentation and part localisation were simultaneously used to encode image content into highly discriminative visual features for the tasks of fine-grained visual categorisation. Zhang et al. [143] proposed a pose-aligned network for deep attribute modelling, which uses poselets for part localisation on the bounding box region and incorporates these semantically aligned part patches into deep convolutional networks for the construction of pose-normalised deep representations. However, such annotation requirements are still demanding in practice, leading some works to consider just utilising local annotations during the training phase, rather than the testing stage. Zhang et al. [144] proposed two pose-normalised descriptors that use deformable part models to localise objects and discriminative parts for both fine-grained recognition and attribute prediction. Further on, Zhang et al. [145] designed a part-based R-CNN, based on bottom-up region proposals and geometric constraints, that is capable of inferring the position

of parts as well as the whole object during testing, without the need for a bounding box. Krause et al. [146] used weak bounding box-level labels to generate partial images with a co-segmentation approach, followed by a dense part alignment across all training images. The aligned details of training images are then transferred test images for prediction of corresponding part positions, and thus the test phase is fully unsupervised at the part level. However, even with such a setup, it is not optimal for most fine-grained recognition tasks, as the local annotations required for the training phase still rely heavily on human involvement, which is expensive and subjective.

Weakly-supervised fine-grained image classification Recently, there has been some emerging work aimed at addressing an interesting research question: how can fine-grained images be classified if they are no detailed local annotation of fine-grained objects in both training and testing stage? For instance, the aforementioned attention mechanism can be utilised as an unsupervised method of identifying discriminating regions without additional local annotations. Xiao et al. [147] proposed a two-level attention-based framework for detecting object parts using only image labels, which is the first work to apply attention mechanisms to the problem of fine-grained recognition. In their architecture, there are two independent deep convolutional networks. The object-level attention module is used to select relevant patches that contain parts of certain objects and filter out the background patches; the part-level attention module is trained based on spectral clustering and then acts as a set of part detectors to localise the discriminated parts. Following the same setup, Fu et al. [148] proposed an attention-based multi-scale recurrent deep network for fine-grained image recognition without local annotations, which is capable of detecting discriminative regions and learning informative region-based features at three different scales in a mutually reinforced manner. Further on, Rodriguez et al. [112] proposed a modular architecture-independent attention module to learn the low-level details and select the most discriminative regions for fine-grained classification. The proposed attention module consisted of three main sub-modules: (i) the attention heads, which defined the most relevant regions, (ii)

the output heads, which generated class probability scores, and (iii) the confidence gates, which outputted a confidence score for each attention head.

Although great progress has been reported, there are still several limitations to this design. First, in the absence of local annotations, it is difficult to ensure that the learned attention maps are consistent for each sample. More importantly, these methods tend to deal with the object parts independently, ignoring their intrinsic dependencies, and thus lack the ability to locate multiple discriminating parts simultaneously. To address this problem, Sun et al. [149] introduced an attention-based framework that automatically detects multiple discriminative parts simultaneously among different input images without requiring part-level labels or explicit cropping of regions of interest. Given a set of training images, they first extracted features from multiple attention regions of each input image, and then applied a multi-attention multi-class constraint to enforce correlations within and between images, i.e., in a metric learning framework, bringing same-attention same-class features closer together and pushing different-attention different-class features farther away. Similarly, Zheng et al, [150] designed a convolutional neural network that is equipped with a multi-attention module to jointly learn multiple attention regions with strong discrimination capabilities. To further exploit the rich part information, a channel grouping loss is introduced for consistent and diverse part model learning, converging the loss function by imposing geometric constraints on the region-based part representations. In addition, some works converted the problem of finding multiple areas of visual discrimination simultaneously into finding them in multiple searches. Zhao et al. [151] proposed a diversified visual attention network (DVAN) for fine-grain object classification, aiming to exploit the computational attention at different scales, especially the small parts or components of fine-grained objects. DVAN first generated a large number of attention canvases of different locations and scales, and used Long-Short-Term Memory (LSTM) to simulate the process of locating multiple discriminative regions simultaneously, as it is able to model long-term dependencies in a recurrent fashion. Zhang et al. [119] extended their ideas by presenting a trilinear attention sampling network (TASN),

which enables the network to learn fine-grained features from hundreds of part proposals. Specifically, a trilinear attention module first generates attention maps by modelling the inter-channel correspondences, and an attention-based sampler randomly highlights one attended part with high resolution in each iteration.

Rather than using attention mechanisms to explicitly detect regions of information in an image, another group of work leverages an unsupervised region proposal method to generate a large number of proposals for discriminative part/object location. For example, Zhang et al. [152] proposed a weakly-supervised fine-grained image categorisation system consisting of region proposal generation and distinctive part selection. After extracting redundant object proposals, they selected the informative patches by clustering the region proposals and exploring the amount of information in each cluster based on importance scores. Finally, the part proposals with the highest importance scores were encoded into a global image representation for the final classification. Yang et al. [153] proposed a two-stage pipeline for fine-grained image classification, which was trained with reinforcement training. Specifically, their Navigator-Teacher-Scrutiniser(NTS) network consists of a Navigator agent, a Teacher network, and a Scrutiniser network. The Navigator generated numerous region proposals for each input image, and produced a list denoting the informativeness of all proposals. The regions with the highest amount of information were cropped and fed into the Teacher network to obtain confidence scores. In the Scrutiniser network, they concatenated all local features for final image-level classification. Recently, Ge et al. [154] provided a different perspective for fine-grained recognition and introduced a new classification pipeline, called the weakly supervised complementary part model, which is based on a region proposal approach that attempts to retrieve local information in a weakly supervised manner from complementary object parts rather than only from the aforementioned most discriminating parts. To capture complementary information, they first applied Mask R-CNN [53] and CRF [155] to generate instance segmentation masks in an iterative manner and selected complementary object proposals based on the segmentation results, followed by feature fusion with two stacked LSTMs [156].

Despite the great performance, the computational cost of processing and selecting candidate region proposals in these methods is extremely high.

2.4 Summary

This chapter discussed several limitations of cell segmentation and biomedical image classification. Section 2.1 first describes both traditional and deep learning-based cell segmentation methods. Regarding traditional methods, data-driven approaches consider only pairs of regions at each step, which is not sufficient in the case of highly ambiguous local content. Model-driven methods offer the opportunity to consider multiple connected segments simultaneously, however, these methods strongly rely on manual initialisation and reliable prior knowledge. On the other hand, deep learning-based methods can produce promising segmentation results, but due to the cost of collecting large-scale biomedical datasets and associated pixel-level annotations, it remains important to develop new techniques to maximise the use of existing small datasets. Therefore, this thesis introduces a hybrid approach equipped with a U-Net and a global probabilistic model for the accurate segmentation of densely packed cells in fluorescent microscopy images. The proposed probabilistic model jointly exploits shape prior and global cell interactions to compensate for the limitations of standard deep networks. Section 2.2 reviewed the literature on biomedical image classification. Single-stage networks extract only global features from reduced low-resolution images, which can lead to a loss of local high-resolution detail. Alternatively, two-stage methods have been introduced to preserve local high-resolution features by dividing the input image into local patches of equal size. However, these pre-extracted patches may ignore long-range correspondences and suppress global structural information such as the density of the object. Based on the recent advances discussed in Sections 2.3, this thesis introduces an attention-based multi-task learning framework for fine-grained classification of biomedical images that is capable of learning complementary fine-grained features from high-resolution imaging data.

3

Global Probabilistic Models for Cell Segmentation

Contents

3.1 Introduction 51
3.2 Deep learning-based Segmentation 54
3.2.1 U-Net Architecture 55
3.2.2 Training Details 57
3.2.3 Experimental Results 58
3.2.4 Failure Cases 59
3.3 Model-based Segmentation 60
3.3.1 Model Construction 61
3.3.2 Algorithm Overview 65
3.3.3 Affinity Functions. 70
3.3.4 Analysis of Segmentation Result 72
3.3.5 Analysis of Global Optimisation 75
3.4 Overall Segmentation 75
3.4.1 Hybrid Segmentation Framework 75
3.4.2 Experimental Results 79
3.5 Conclusions 80

This chapter presents a hybrid approach for robust and reliable delineation of individual objects in a densely packed population of cells. The core of this pipeline is to combine DCNN-based segmentation approaches with a global probabilistic model that explicitly enforces prior shape constraints and global cellular interactions.

While deep learning has dramatically improved our capabilities for developing robust segmentation methods, the relationship between algorithm performance and the required amount of training data is not well understood. This chapter offers a preliminary and fundamental step in analysing small biomedical datasets and incorporating global context information.

Originality: The high-throughput cellular imaging data was provided by collaborator Dr. Chandan Seth Nanda. I manually labelled cell nuclei on 50 images as ground truth for the segmentation task. The main principle of the proposed probabilistic model is derived from the approach of Rittscher et al. [157]. I independently implemented this algorithm on cell images using Python with the following extensions: (1) The model's parameter space is extended to cover a comprehensive range of region-based features, including shape, texture and statistical information. (2) I designed a parametric elliptical model for the cell images to explicitly enforce the shape priors. (3) I performed edge analysis on each input image to speed up the runtime and improve the segmentation accuracy. I also proposed the idea of combining the global probabilistic model with a U-Net and demonstrated that this hybrid approach can effectively handle both independent cells and touching cells. The main ideas presented here were first presented in:

Mengran Fan and Jens Rittscher. Global probabilistic models for enhancing segmentation with convolutional networks. *Proceedings of the International Symposium on Biomedical Imaging (ISBI)*, 2018. - Oral presentation.

3.1 Introduction

Fluorescent microscopy imaging is a sensitive and quantitative method that is widely used for a variety of target discovery and quantitative cellular studies. With the help of multi-channel fluorescent labelling, different fluorescent dyes are used to label specific cellular structures and therefore the molecular components can be identified separately in the same sample. The standard image acquisition process for multi-channel fluorescent imaging is illustrated in Figure 3.1.

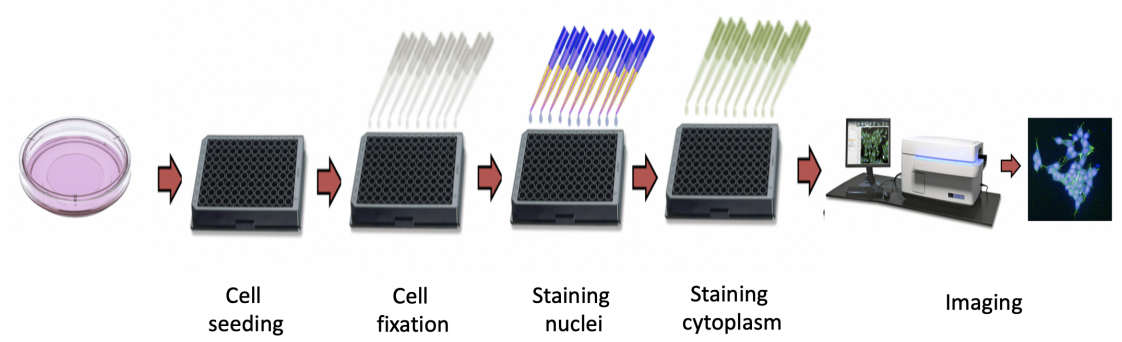


Figure 3.1: The standard image acquisition process for multi-channel fluorescent imaging.

However, while certain sample pre-processing methods can be utilised to control the overall plating density of cells, it is still very difficult to avoid their aggregation in realistic experimental settings. Moreover, as the field progresses towards developing more complex in vitro cell models that promise to be more phenotypically relevant and predictive, there is an increasing need for reliable automatic delineation of densely packed cellular objects. To address this, this chapter investigates a globally optimal model-based approach for the automatic segmentation of cell nuclei in fluorescent microscopy images.

In summary, most of the challenges associated with cellular imaging data are rooted in visual complexity, high variability in cellular nuclear morphology, and ambiguity associated with dense clustering of cell nuclei. The existence of intensity inhomogeneities and imaging artefacts, such as water droplets, bubbles or dust particles, usually lead to undesirable image features (see Figure 3.2 (a)). Meanwhile, It can also be observed that the staining of cell nuclei lacks specificity. After splitting channels, parts of the cytoplasm may also be highlighted in the fluorescent channel for cell nuclei (shown in Figure 3.2 (b)). Besides, cell nuclei typically appear in the microscopy images as densely packed or partially overlapped objects (shown in Figure 3.2 (c)). Therefore, designing a method for delineating individual cell nuclei from the background of fluorescence microscopy images that could function well in practical settings is complex and requires deliberate design.

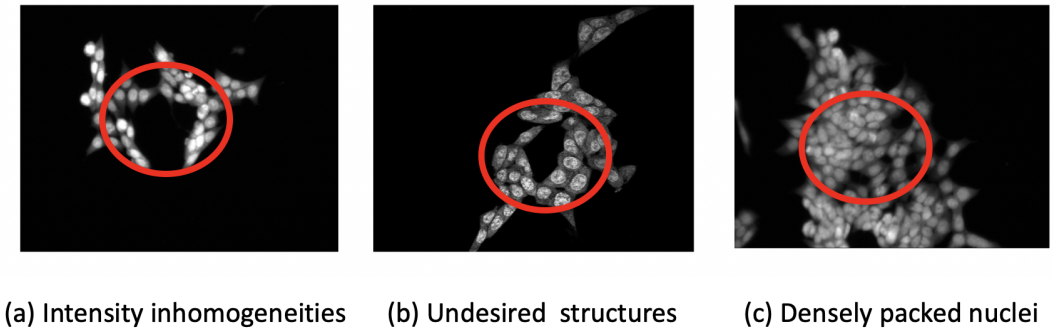


Figure 3.2: Illustration of the challenges in automatic segmentation of cell nuclei.

Section 2.1 provides a comprehensive review of cell segmentation. Most of the early cell segmentation methods were based on data-driven and model-driven techniques. Data-driven methods [13, 15, 16] focus on local image feature grouping, but only consider pairs of regions at each step. Model-driven approaches [45–47] offer an opportunity to give global control for multiple segments, but strongly rely on reliable prior knowledge. Recently, deep learning-based approaches [56, 57] have achieved significant success in biomedical image segmentation. However, in practice, well-annotated biomedical imaging data may not always be available. Acquiring relevant pixel-level ground truth is also expensive and often limited by the available expert resources. To address the challenge of limited data annotation, there is growing interest in developing segmentation methods that require only a small amount of annotated data, such as transfer learning, semi-supervised learning or self-supervised learning. Here, we propose a hybrid segmentation framework where a global probabilistic model is used to compensate for the failure case of U-Net. The motivation for the global probabilistic model is to capture the advantages of data-driven and model-driven techniques by incorporating global shape prior information into local feature grouping methods. In this way, shape prior information is explicitly incorporated when inferring the label of each segment, and the global optimisation process is not affected by local minima.

In the remainder of this chapter, the baseline performance of U-Net and some failure cases are first discussed to motivate the design of our segmentation framework (see Section 3.2). The proposed global probabilistic model (see Section 3.3) and a hybrid segmentation framework are also presented to enhance the segmentation performance on a small dataset.

3.2 Deep learning-based Segmentation

Deep neural networks (DNNs) have recently been shown to excel in a wide range of biomedical image segmentation tasks because of their ability to automatically learn comprehensive feature representations from data, rather than traditional hand-crafted features. In [52], Long et al. proposed the groundbreaking Fully Convolutional Network (FCN) for pixel-wise prediction, dramatically improving the state-of-the-art. However, the deconvolutional layer in this architecture is simply implemented by a fixed bilinear interpolation. Therefore, the absence of learning-based deconvolution makes it difficult to capture accurate boundaries for individual objects. To resolve such issue, Noh st al. [158], proposed a multi-layer deconvolution network featured with multiple series of unspooling, deconvolution and recitation layers. The learnable deconvolution network tends to generate accurate pixel-level segmentation masks and enables the accurate reconstruction of the detailed object structure. Inspired by deep deconvolution networks, Ronneberger et al. [5] proposed an encoder-decoder architecture with skip connections that transfer fine-grained features from the encoder sub-network to the decoder sub-network along with matching levels, achieving the state-of-the-art performance in various biomedical image segmentation tasks [56, 58]. This section explores the deep learning-based methods that generally follow the architecture of the original U-Net [5]. It is aimed to investigate the performance of U-Nets in automatic segmentation of densely packed objects, and thereby motivate how a model-based framework can be incorporated to improve the performance when only limited training data is available.

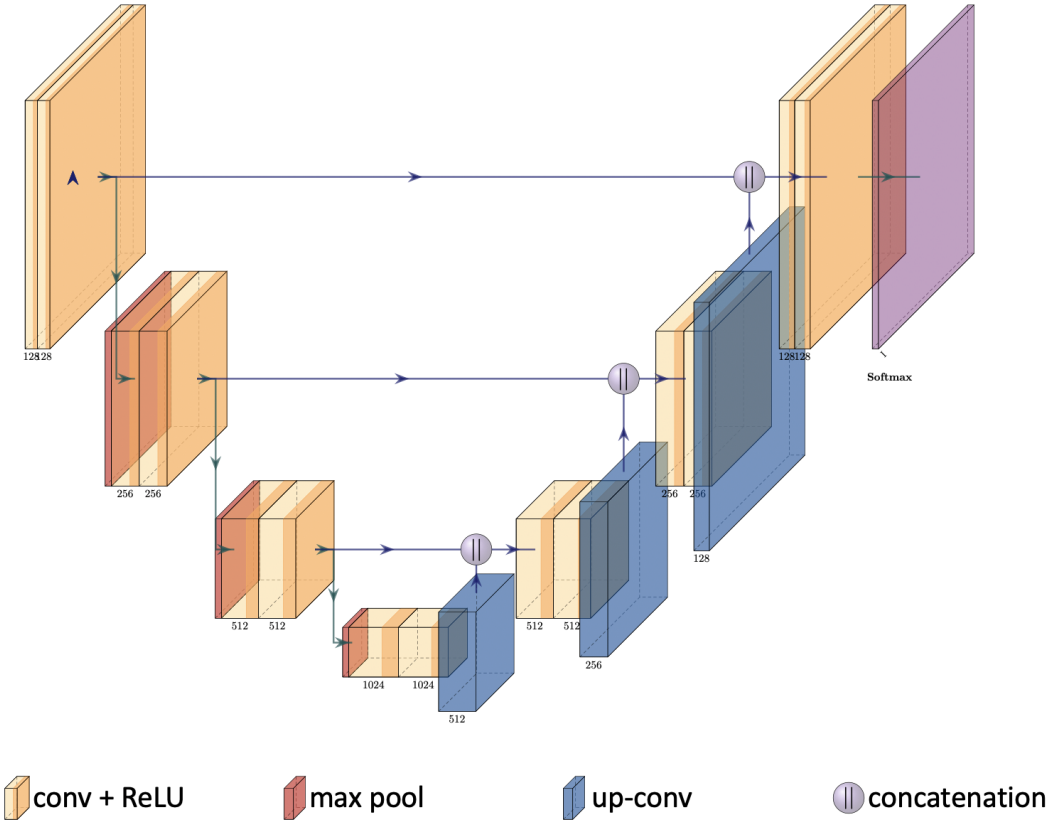


Figure 3.3: U-Net architecture with three stages and 128 channels in the first layer.

3.2.1 U-Net Architecture

Ronneberger et al. [5] proposed an elegant architecture that can be trained end-to-end from very little training data available and achieved state-of-the-art performance on cellular object segmentation. As shown in Figure 3.3, the overall architecture is composed of two parts: encoding path and decoding path. The encoding path corresponds to feature extractor that transforms the input image to a low-dimensional representation, whereas the decoding paths is an image generator that expands the extracted feature maps back to the original size. The additional skip connections transform the fine-grained details from the shallow layers in the encoding branch to the deep layers in the decoding branch along the matching levels. The final output of the network is a probability map in the same size to the input image, representing the probability that each pixel belongs to a pre-defined class.

This success is largely due to the skip connections in the architecture, which allows the network to consider both shallow, low-level feature maps in the contraction path and deep, semantic feature maps in the symmetric extension path. As the architecture has been proven to be effective in recovering multi-scale fine-grained details of cellular objects even in complex contexts, and outperforms previous best methods on multiple cellular datasets, this chapter will use U-Net as an entry point to investigate the performance of deep learning-based cellular segmentation methods. More importantly, their original paper introduced a weighted loss to solve the segmentation problem of touching cells, where pixels in the border regions receive a large weight in the loss function. Therefore, this strategy was also investigated in this chapter to show whether the weighted loss scheme can facilitate accurate segmentation of dense cell nuclei.

I first build and modify this architecture to make it work for my small dataset. In the encoding path, max pooling is designed to sequentially eliminate the confounding neuron activations in a lower layer by summarising the most activated features in a receptive field. These pooling layers both greatly reduce the number of model parameters and achieve a degree of invariance to the local translation of the input. Consequently, the higher layers of the network increasingly retain only abstract representations of the input image, which plays an important role for most high-level tasks (i.e., classification and object detection). Despite the robustness, the key downside associated with this design is that the spatial information within a receptive field is lost at higher layers due to the significant downsampling in the encoder path, which limits the performance of deep learning-based methods in precise pixel-level segmentation tasks. To compensate for the loss of spatial details, the number of feature maps from the first layer in the encoding path was increased (128 channels in the first layer and double the number of channels at each step). Besides, the convolution transposition layers were employed to reconstruct the spatial dimensions of the feature maps in the decoding path. This up-sampling strategy is similar to the original design of U-Nnet [5], but does not face the cropping issues due to the complete symmetry between the encoding and

decoding paths. Compared with the original U-Net [5], this design also reduces the number of max pooling operations to build a relatively shallow network, and in this way too much spatial information loss is avoided. Finally, the adaptive weighting schemes were incorporated in which a larege weight was assigned to edge pixels between touching cells in the loss function. Unlike previous binary classifiers that only discriminate nuclei against the background, the segmentation models with a weighted loss function can predict the inner region of the nuclear, the nuclear contour and the background at the same time.

3.2.2 Training Details

Dataset. In collaboration with Sebastian Nijman's lab at the Target Discovery Institute at the University of Oxford, Dr. Chandan Seth Nanda provided me with high-throughput imaging data of haploid cells, where DAPI was used to mark the cell nuclei. In my experiments, the cellular microscopy image dataset consisted of 30 images for training and 20 images for testing. In the training phase, 128 x 128 patches were randomly extracted from 30 annotated images of 1024 x 1360 pixels. Since the training data available in our task was very limited, the data augmentation techniques were used to avoid over-fitting. Small rotations, horizontal flips and contrast distortions on the available training images were performed to make the model more robust to the location and intensity of various cells. Before training, each patch needs to be normalised by subtracting the average of all images in the training set and then dividing by the standard deviation of the entire training set.

The proposed network was implemented in tensorflow with Keras library. In total, the model contains about 4.5 million trainable parameters, which are initialised with an orthogonal basis. In addition, the parameters of the network are optimised by stochastic gradient descent (SGD) with a momentum of 0.9. The initial learning rate of the SGD and weight decay are set to 0.001 and 0.0005, respectively. The network converges after approximately 50 epochs with mini-batch of 16 samples.

In our experiments, two U-Net models, U-Net-A and U-Net-B, were trained separately, in which the cross-entropy loss was computed for each pixel in the

generated probability map. However, in U-Net-B, I increased the weights of pixels belonging to the boundary class to force the network to focus on the small boundaries between densely packed cells. Following [5], a pixel-level weight map was pre-defined for each ground truth segmentation to deal with the class imbalance in the training set.

Concretely, as in [5], the loss function of U-Net-A and U-Net-B are formally given below in Eq. 3.1 and Eq. 3.2, respectively.

$$L_{net-A}(X_0) = -\sum_{i}^{n}(P_{i,y_{true}} \log(P_{i,y_{pred}}) \tag{3.1}$$

$$L_{net-B}(X_0) = -\sum_{i}^{n}(P_{i,y_{true}} \times w_i \times \log(P_{i,y_{pred}}) \tag{3.2}$$

Here, n represents the total number of classes in this dataset, $P_{i,y_{true}}$ is the ground-truth probability map with respect to the input patch X_0 on the i_{th} class, $P_{i,y_{pred}}$ is the output probability of the network, and w_i is the weight value pre-defined to provide more importance to the boundary pixels during training.

3.2.3 Experimental Results

The U-Net-A and U-Net-B were applied to 20 test images containing approximately 1500 cell nuclei. To study the segmentation accuracy, I collected the predictions L_{pred} of two U-Net models and calculate two overlap radio measures based on the ground truth segmentation masks L_{true}.

Jaccard index. Jaccard index, also known as Intersection-Over-Union (IoU), is a metric commonly used in segmentation masks, which indicates the area of overlap between the prediction and ground truth divided by the area of union between the predicted result and ground truth. The measure is defined as:

$$Jaccard(L_{pred}, L_{true}) = \frac{\left|L_{pred} \cap L_{true}\right|}{\left|L_{pred} \cup L_{true}\right|}. \tag{3.3}$$

Dice similarity coefficient. The Dice coefficient is similar to Jaccard index, indicating the area of overlap between the foreground of the predicted result and the foreground of the ground truth. The measure is defined as:

$$Dice(L_{pred}, L_{true}) = \frac{2 \times \left| L_{pred} \cap L_{true} \right|}{\left| L_{pred} \right| + |L_{true}|}. \tag{3.4}$$

The Jaccard index and Dice coefficient both range from 0 to 1, with 1 indicates complete congruence, and 0 means no overlap. The quantitive results of U-Net-A and U-Net-B are summarised in Table 3.1.

Method	Jaccard Index	Dice Index
U-Net-A	0.67	0.80
U-Net-B	0.72	0.84

Table 3.1: Segmentation performance of U-Net-A and U-Net-B. The U-Net with a weighted loss function (U-Net-B) gave higher Jaccard and Dice scores for the segmentation of dense cell nuclei, demonstrating the effectiveness of incorporating a weighted training scheme.

3.2.4 Failure Cases

The examples of failure cases of U-Net-A and U-Net-B are illustrated in Figure 3.4. It can be concluded that U-Net-A can effectively capture the isolated nuclei, whereas cell clumps are typically recognised as single objects. Alternatively, U-Net-B tended to separate the cluttering nuclei successfully by adopting the weighted loss function, whereas most pixels in the interior and the cytoplasm neighbouring the boundaries were also predicted as boundaries in the testing phase, leading to low pixel-level accuracy. Therefore, although this adaptive weighting scheme improves and stabilises the inner nuclei and boundary prediction, it is highly sensitive to the pre-defined weights in the weighted loss function. Although one could make use of additional training data or transfer learning to overcome this problem, it is important to design robust algorithms on small biomedical datasets. This thesis will demonstrate later that the proposed global probabilistic model, which is introduced in the next section, can be used to overcome these failure modes.

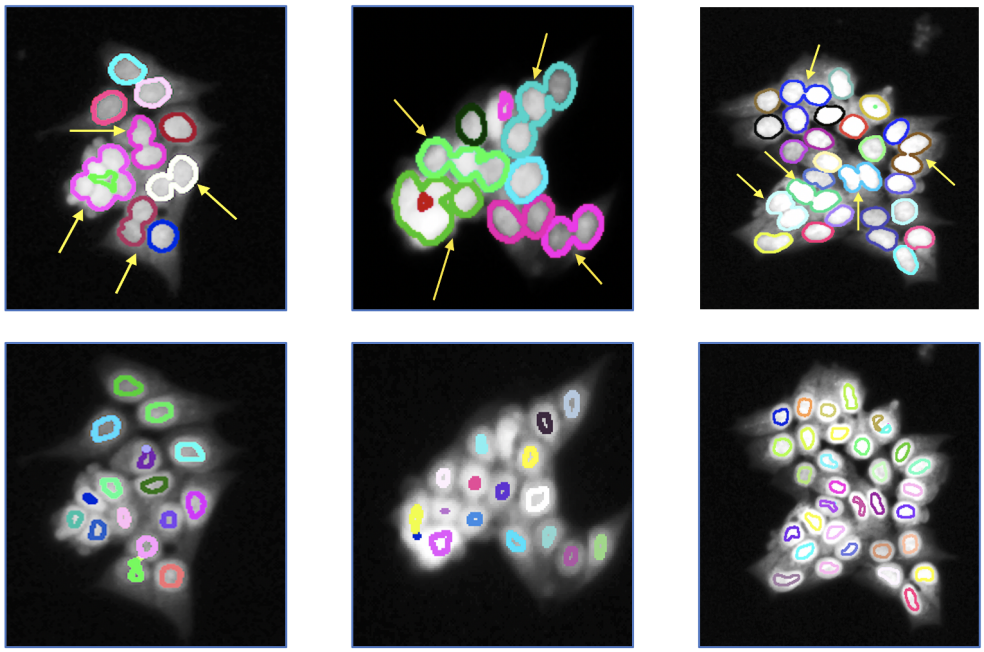


Figure 3.4: Failure modes of U-Net based segmentation. The first row presents the result of U-Net-A which is trained using an equally weighted loss function. It can be observed that the network identifies a clump of cells as one object. The second row shows the failure modes of U-Net-B with a weighted loss function. Although it can separate the clustered nuclei successfully, the delineation of the object boundaries is not very accurate.

3.3 Model-based Segmentation

This section describes the proposed global probabilistic model that can be used as a final step to effectively compensate for the shortcomings of the segmentation obtained by a CNN-based method, as the examples shown in Figure 3.5. The main focus of this work is to address the segmentation of densely packed objects by jointly exploiting shape priors and global contexts. In the fluorescent channel of cell nuclei, it can be seen that the cell nuclei are typically bright and have a shape that generally approximates an ellipse. Their contours can be explicitly represented by a parameterised elliptical model. Moreover, it is desirable to design a globally optimal approach to better cope with the loss of spatial details in deep learning-based methods. Thus, I propose a globally optimal model-based segmentation method, which explicitly captures the shape priors and groups the features in a

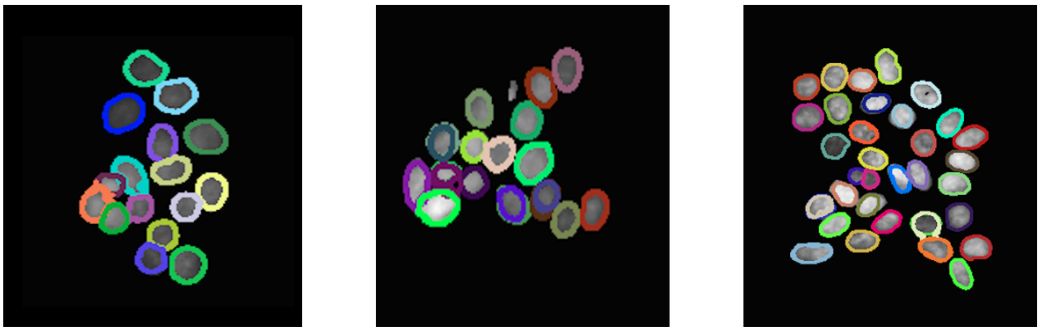


Figure 3.5: Model-based segmentation. The figure shows selected results of our model-based segmentation methods that correspond to failure modes that were highlighted in Figure 3.4.

globally consistent manner. The algorithm is based on an existing framework for the segmentation of groups of people into individuals [157], where the core principle is to assign local image observations to a set of target object hypotheses which are associated with a prior shape model.

3.3.1 Model Construction

The set of image observations $Z = \{z_i\}$ consists of various region-based image features, such as edges, intensities, gradients, based on an initial superpixel step [27]. Which subsets $H \subset Z$ should be grouped together to a single cell is identified by a generative model that jointly uses shape and intensity information. As an initial estimate, each collected superpixel segment is considered as a single cell hypothesis h. The set of all hypotheses is defined as:

$$H := \{h_1, h_2, h_3, ..., h_K\} \tag{3.5}$$

The collection of model parameters x_i that encodes shape, intensity and location information of each potential hypothesis h_i is denoted as:

$$X = [x_1, x_2, x_3, ..., x_K]^T \tag{3.6}$$

With the local image features and initial hypotheses, the assignment vector V of length K is directly defined to associate each image segment z_i with a particular

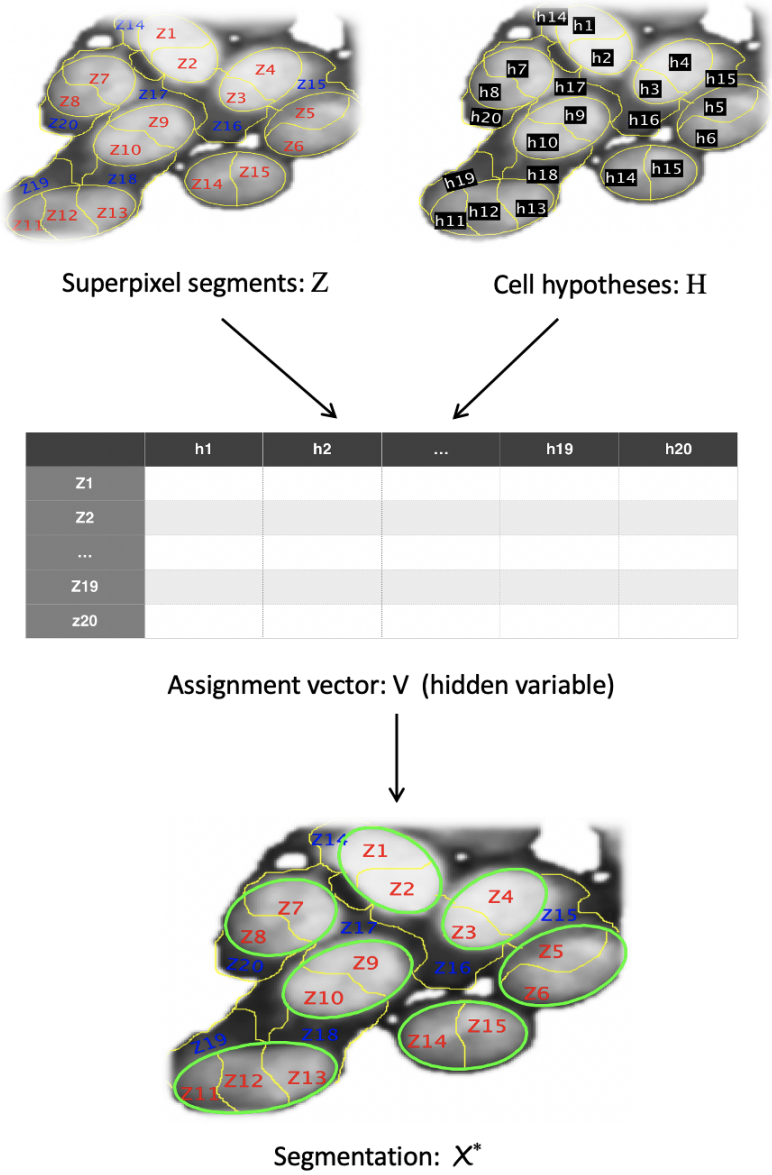


Figure 3.6: Illustration of the proposed model-based approach. The core principle of the proposed segmentation model is to assign image observations Z to a set of object hypotheses H that are associated with model parameters X. The set of image observations Z consists of various region-based image features based on an initial superpixel step. As an initial estimate, each collected superpixel segment is considered as a single cell hypothesis h_i. With the local image features and initial hypotheses, the assignment vector V is directly defined to associate each image segment z_i with a particular hypothesis h_k. It is worth noting that the assignment vector V is treated as a hidden variable since the assignment of segment z_i to hypothesis h_k can not be observed directly. The segmentation on the bottom shows the maximum a posteriori (MAP) estimate of the model parameters X^*.

hypothesis h_k. It is worth noting that the assignment vector V is treated as a hidden variable since the assignment of segment z_i to hypothesis h_k can not be observed directly.

Therefore, the focus of this method is to model the joint probability of the assignment vector V and the image feature set Z, defined as $P(V, Z; X)$. The important part of this approach is that maximum likelihood estimates of both the model parameters X and the unobserved assignments V can be obtained simultaneously. Prior work motivates that we can estimate X by optimising the log likelihood $L(V|Z; X)$ of a feature assignment V given a set of image features Z. It is important to note here that this approach is an effective global optimisation that takes into account all the information of the image region.

The joint probability model $P(V, Z; X)$ is based on a set of merit functions that measure the affinity of a single or a pair of image features with a given hypothesis. Here I mainly capture the affinities of single feature assignments as well as pairwise assignments to a given potential hypothesis h_k with parameter x_k. The corresponding affinity matrix is defined as $g(x_i, h_k)$ and $g(x_i, x_j, h_k)$. In practice, the feature assignment V is represented as a matrix $M = M_{ik}$ where m_{ik} is the probability of i^{th} image feature $z_i$belonging to the k^{th} hypothesis h_k. When the affinity functions and assignment matrix is incorporated, the log likelihood of a feature assignment V given a set of image feature Z is defined as follows:

$$\begin{aligned} L(V|Z; X) \propto & \gamma_1 \sum_{k=1}^{K} \sum_{i=1}^{K} g(z_i, x_k) m_{ik} + \\ & \gamma_2 \sum_{k=1}^{K} \sum_{i,j=1; i \neq j}^{K} g(z_i, z_j, x_k) m_{ik} m_{jk} \end{aligned} \tag{3.7}$$

Note that the value of $p(Z)$ remains constant throughout the formulation, the joint probability shall be reformulated as:

$$p(V, Z; X) = p(V|Z; X)p(Z) \propto \exp(L(V|Z; X)). \tag{3.8}$$

The method of Rittscher et al. [157] demonstrates that the maximum a posteriori (MAP) estimate of X can be found by a variant of the Expectation-Maximisation

(EM) algorithm, which iteratively maximises the free energy function in both E-step and M-step. The energy function $F(\tilde{P}, X)$ is defined as:

$$F(\tilde{P}, X) = E_{\tilde{P}}[\log P(V, Z; X)] + H(\tilde{P}), \tag{3.9}$$

where $\tilde{P}$ is the probability density function of$V$, and H denotes the entropy of this distribution. Each iteration in the EM algorithm consists of two main steps: the expectation step (E-step) and the minimisation step (M-step). The E-step finds the optimal distribution of V under the current model parameters, and the M-step updates the model parameters X to increase the energy function while assuming that the probability distribution found in E-step is correct. The process starts with a random initial guess and ends until it converges. An iteration of the EM algorithm can be expressed as follows:

$$\begin{aligned} &\text{E-step: Set } \tilde{P} \text{ to the } \tilde{P} \text{ that maximises } F(\tilde{P}, X) \\ &\text{M-step: Set } X^* \text{ to the } X \text{ that maximises } F(\tilde{P}, X). \end{aligned} \tag{3.10}$$

In order to simplify the energy function, I assume the $\tilde{p}$ is independent with respect to the v_i, i.e.

$$\tilde{P}(V) = \sum_{i=0}^{N} \tilde{P}(v_i) \tag{3.11}$$

Now based on formulation specified in Eq. 3.7 and Eq. 3.8, the energy function $F(\tilde{P}, X)$ is of the form:

$$\begin{aligned} F(\tilde{P}, x) = &\sum_{k=1}^{K}\sum_{i=1}^{K} g(z_i, x_k) m_{ik} + \sum_{k=1}^{K}\sum_{i,j=1;i\neq j}^{K} g(z_i, z_j, x_k) m_{ik} m_{jk} \\ &- \sum_{k=1}^{K}\sum_{i=1}^{K} m_{ik} \log(m_{ik}) \end{aligned} \tag{3.12}$$

The optimal distribution of V that maximise the joint probability $P(V, Z; X)$ can be defined as v^*,i.e.

$$P(V^*, Z; X) \geq P(V, Z; X), \forall V \in \mathcal{V} \tag{3.13}$$

It can be seen that:

$$\forall \tilde{P} \log P(V^*, Z; X) \geq E_{\tilde{P}}[\log P(V, Z; X)] \tag{3.14}$$

Therefore the expectation term in the energy function $F(\tilde{P}, X)$ is maximal when $\tilde{P}(V) = \delta(V^*)$. On the other hand, The entropy term favours broader distributions, i.e., the greater the number of distributed hypotheses, the greater the entropy of the probability distribution $\tilde{P}$. Inspired by [157], An explicit regularisation parameter $T$is introduced to make this optimisation problem more tractable:

$$F(\tilde{P}, X) = E_{\tilde{P}}[\log P(V, Z; X)] + TH(\tilde{P}), \tag{3.15}$$

The temperature parameter is aimed to globally control how broadly a set of probabilities is distributed. When the starting temperature is sufficiently high $\tilde{P}$ will be a uniform distribution over a variety of hypotheses, since the value of the energy function F is dominated by the entropy term. This is followed by an annealing process, which gradually reduces the temperature, and the probability distribution thus becomes narrow.

3.3.2 Algorithm Overview

Image regions as components of the object model. Image regions (superpixels) generated from locally coherent groupings of image pixels provide a more natural representation of image data compared to individual pixels. This work leverages a popular superpixel segmentation algorithm, simple linear iterative clustering (SLIC) [27], to facilitate the extraction of perceptually meaningful segments from images. SLIC adapts a k-means clustering method to efficiently generate superpixels and provides flexibility in the compactness and number of the superpixels, which is shown to yield state-of-the-art adherence to image boundaries, as depicted in Figure 3.7.

Capture relevant image features. For each image segment, a region-based feature descriptor is constructed to describe the intensity, shape, location and texture information. This descriptor provides the 17 relevant features that are derived from location, intensity and texture information. Table 3.2 gives a complete summary of different features collected from the segment.

Image feature	Type	Feature description
Area	Geometrical	Actual number of pixels in this segment
Perimeter	Geometrical	Actual number of pixels on the boundary of the segment
Centre	Geometrical	Coordinates of the segment centroid
Major axis length	Geometrical	Length of the major axis of the ellipse that best fits the contour
Minor axis length	Geometrical	Length of the minor axis of the ellipse that best fits the contour
Circularity	Geometrical	Degree of roundness of a circular object
Fitness	Geometrical	Ratio of the area between the segment and the bounding ellipse
Mode	Statistical	The most frequent value of the pixels in the segment
Mean	Statistical	The average value of the pixels in the segment
Standard deviation	Statistical	Standard deviation of the pixels in the segment
Variance	Statistical	Variance of the pixels in the segment
Sum	Statistical	Sum of the pixels in the segment
Homogeneity	Texture	Intensity variation of the segment
Contrast	Texture	Intensity contrast between neighbouring pixels
Correlation	Texture	Intensity correlation/stability in the segment
Entropy	Texture	Intensity smoothness of the segment
LBP histogram	Texture	Intensity distribution in the segment

Table 3.2: Representative image features for the cell description. For each image segment, a region-based feature descriptor is constructed to describe the geometrical, statistical and texture information. This table provides a complete summary of the 17 relevant region-based features. Note that the aim of this study is to demonstrate the principles of this segmentation model. In a straightforward manner, I have chosen a set of common features used in data-driven cell segmentation methods. The type and number of features can be further adapted for different tasks.

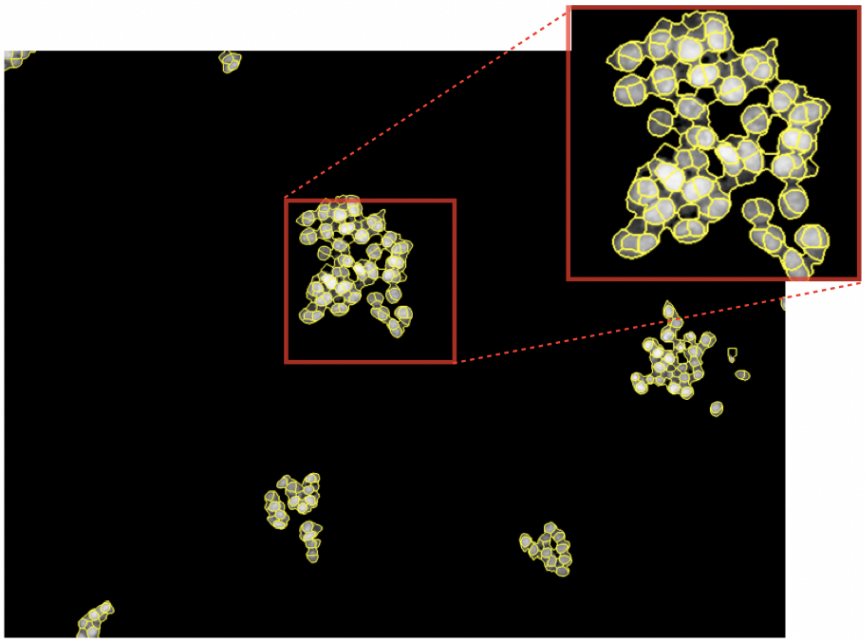


Figure 3.7: Sample result of the SLIC superpixel method. The SLIC algorithm [27] is widely used to efficiently generate locally consistent patches. It is clear that the over-segmented regions adhere well to the boundaries of the cell nuclei.

Hypothesis shape model. The cell nuclei, despite its different size and brightness, can be closely approximated mathematically as an ellipse. This provides some intuition to define a parameterised elliptical model as a shape prior. In practice, the ellipse that best fits the contour of the image segment can be extracted by the least square fitting algorithm, parameterised as $S = (x, y, m, n, \Theta)$. The coordinates of the centroid (x, y) is obtained by finding the mean of pixel position along the x and y axis. The lengths of the major and minor axes are denoted by m and h. The orientation of the ellipse (the angle between the major axis and the image axis) is described by angle Θ.

Extended parametric space. To maintain the generality of the algorithm, the model parameters also encode the texture pattern and intensity variance for each hypothesis. Specifically, the texture descriptor is efficiently captured by the local binary pattern (LBP) operator, which computes a normalised texture histogram $T_i = \{t_i^1, t_i^2, ..., t_i^n\}$ to represent the stability of the intensity and its distribution

among n bins. In addition, a statistical summary of pixel intensities is derived from intensity and position information, i.e., maximum, minimum, and mean intensity.

Estimation of model parameters Instead of traditional EM algorithm, a variation version is adopted to this framework to find the maximum likelihood estimate of both model parameters and feature assignments. In this version, E-step and M-step both iteratively maximise, or at least increase the function $F(\tilde{P}, X)$ defined above. For E-step, it is aimed to maximise the $F(\tilde{P}, X)$ with respect to the assignment probabilities m_{ik} in matrix M. For every $i \in [1, ..., K]$ and $k \in [1, ..., K]$, the assignment probability m_{ik} is updated as:

$$m_{ik}^{'} = \exp(\frac{1}{T}\partial_{m_{ik}} E_{\tilde{P}}[\log P(V, Z; X)]), \tag{3.16}$$

where the partial derivative $\partial_{m_{ik}} E_{\tilde{P}}[\log P(V, Z; X)])$ has the form:

$$g(z_i, x_k) + 2 \sum_{j=1, i \neq i}^{K} g(z_i, z_j, x_k) m_{ik} \tag{3.17}$$

Based on the independent assumption described above, the matrix `M`can be normalised as follows:

$$m_{ik}^{s+1} = \frac{m_{ik}^{'}}{\sum_l m_{il}^{'}} \tag{3.18}$$

In the M step, the model parameters X are updated in the presence of the current assignment probability matrix M. In the experiments, due to the relatively small number of candidate hypotheses and model parameters, I simply use linear combinations and exhaustive search to achieve this step, i.e., the weaker a particular assignment probability m_{ik} is, the less it affects the shape parameter x_k. An overview of the proposed algorithm and the preliminary result of one sample image is shown in Algorithm 1 and Figure 3.8, respectively.

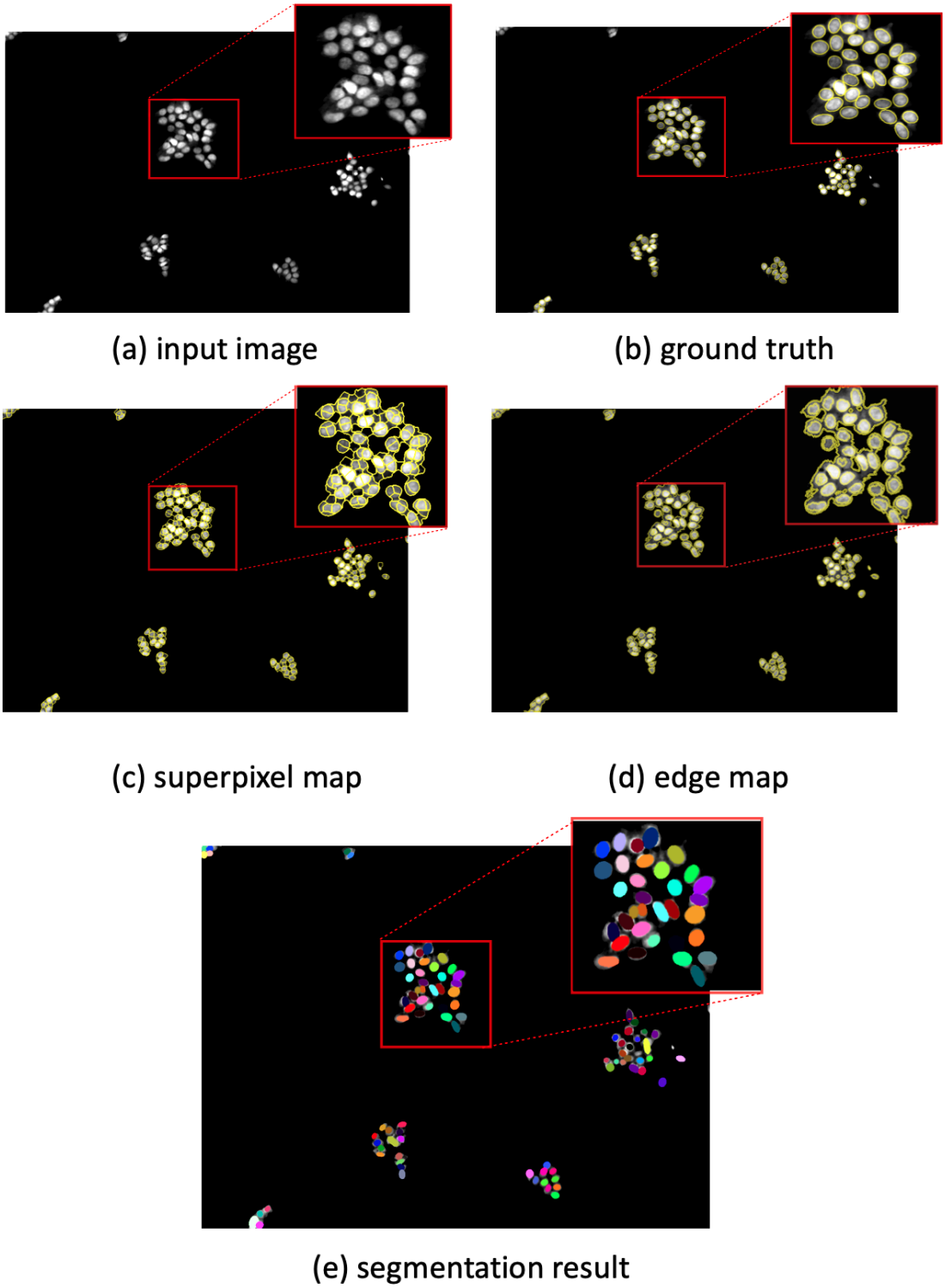


Figure 3.8: Preliminary result for one sample image. (a) Input image. (b) Ground truth. (c) Over-segmented result generated by SLIC superpixel algorithm. (d) Resulting edge map of canny edge detector. (e) Segmentation result of the proposed model.

Algorithm 1: Overview of the algorithm

Pre-processing;
- Run superpixel method and edge detector to generate image segments and edge map;

Initialisation;
- extract region-based image features;
- initialise model parameters;
- set T = T_0

while *not at point of convergence* **do**
  E-step;
  - Update $T = \alpha T$;
  - Update the affinity matrix;
  - Update the probabilities in assignment matrix M;

  M-step;
  - Update the model parameters with the current likelihood distribution;

end

Hypotheses validation;
- Detect the cell nuclei objects from the segmentation results;

Downstream analysis;
- Summarise model parameters and region-based attributes;
- Export to a csv file;

3.3.3 Affinity Functions.

To make it robust to various image conditions, the joint likelihood model encodes single feature assignments as well as pairwise assignments for a given candidate object hypothesis. Each hypothesis specifies an ellipse with a rich set of region-based feature descriptors. Given a single image segment, the distance, texture and intensity features are used to assess how well it fits a particular hypothesis. Given a pair of segments, texture and edge information are typically utilised to measure whether they should be merged. For example, if two segments have very similar internal texture, it is likely that they belong to the same candidate object hypothesis. In case two segments are divided by a strong edge it is unlikely that they belong to the same cell. The parameters of the affinity functions are optimised on a set of sample image patches. Pilot experiments showed that the selection of parameters did not significantly affect the segmentation results, but did increase the processing time. Each affinity function is defined as follows.

Image ID	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	total
Number of nuclei	98	149	90	86	137	40	82	78	121	73	954
Correctly segmented	89	133	84	81	124	32	72	69	105	63	852
Under-segmented	1	5	3	4	1	2	1	2	0	3	22
Over-segmented	3	4	0	0	5	5	5	0	1	2	25
Boundary errors	2	3	1	0	4	0	3	1	3	4	21
Missed	3	4	2	1	3	1	1	6	8	1	30
Precision	0.95	0.94	0.95	0.96	0.89	0.78	0.83	0.95	0.92	0.88	0.91
Recall	0.91	0.89	0.93	0.94	0.91	0.80	0.88	0.88	0.87	0.86	0.89
F1-score	0.93	0.92	0.94	0.95	0.90	0.79	0.85	0.91	0.90	0.87	0.90

Table 3.3: Quantitative summary of 10 sample images.

Distance affinity f_d. Distance affinity is defined based on the Euclidean distance between the region centroids. Let l_n and l_m be the perimeters of regions r_n and r_m, and l_{nm} denotes the distance between their centroids. The distance affinity is defined as:$f_d(n,m) = 1 - \frac{l_{nm}}{l_n+l_m}$.

Texture affinity f_t. For image segments r_n and r_m, I constructed one-dimensional LBP histograms h_n and h_m normalised by the L1 norm, respectively. The texture affinity is measured by the intersection of histograms: $f_t(n,m) = |h_n \bigcap h_m|$.

Intensity affinity f_i. In addition to the intensity distribution, a statistical summary of pixel intensity was generated for each segment, containing mode, mean, standard derivative, and intensity variance. Given the feature vector v_n and v_m, the intensity affinity is calculated by the intersection of vectors: $f_i(n,m) = |v_n \bigcap v_m|$.

Edge affinity f_e. With the edge map generated by Canny edge detector [159], edge information is taken into account to measure whether there are edge points between adjacent segments. To achieve that, a bounding rectangle is created around the line between the region centroids, as shown in Figure 3.9. The height h and the width w can be adjusted based on the length of major and minor axes of r_n and r_m. The edge affinity f_e is determined by the ratio of the number of edge points lied in this bounding box, x, and the sum of height h and width w: $f_e(n,m) = \frac{1-x}{h+w}$

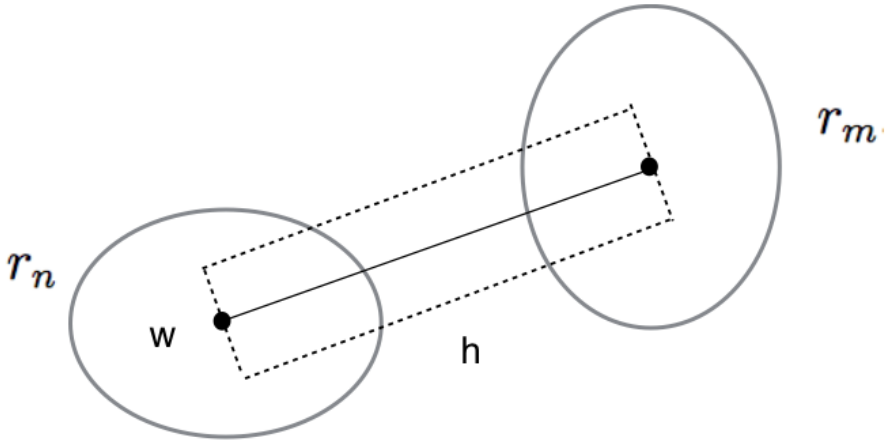


Figure 3.9: Bounding box around the central line of region r_n and r_m. The edge affinity is defined by the ratio of the number of edge points lied in this bounding box and the sum of height and width of this bounding box.

Assignment affinity functions $g(z_i, x_k)$ and $g(z_i, z_j, x_k)$. The final affinity function assigning z_i to hypothesis h_K with parameters x_k is a linear combination of the above affinity measurements. It should be noted that in this implementation, two thresholds are set for distance and edge affinities, i.e., if two regions are far away from each other or there is a strong edge between segments, the final affinity is set to zero directly. The single affinity $g(z_i, x_k)$ and pairwise affinity $g(z_i, z_j, x_k)$ are defined as:

$$g(z_i, x_k) = f_d(z_i, x_k) + f_t(z_i, x_k) + f_i(z_i, x_k)$$

$$g(z_i, z_j, x_k) = f_e(z_i, x_k) + f_e(z_j, x_k) + f_e(z_i, z_j) + f_t(z_j, x_k) + f_t(z_i, z_j)$$

3.3.4 Analysis of Segmentation Result

This section presents experiments on the qualitative evaluation of segmentation accuracy. I evaluated the performance of the model-based segmentation method on 10 cell microscopy images containing a total of 954 nuclei. Object-level segmentation performance is reported in terms of precision, recall, and F1-score. Given the number of true positives (TP, i.e. the cell nuclei that have been correctly captured), true

negatives (TN), false positives (FP) and false negatives (FN), the precision, which describes the portion of positive identifications is actually correct, is defined as:

$$Precision = |TP|/(|TP| + |FP|). \tag{3.19}$$

The recall, which describes the proportion of actual positives was identified correctly, is defined as:

$$Recall = |TP|/(|TP| + |FN|). \tag{3.20}$$

The F1-score, which is the weighted average of Precision and Recall, is defined as:

$$F = 2 \times (Recall \times Precision)/(Recall + Precision). \tag{3.21}$$

Table 3.3 provides a complete quantitative summary of the 10 sample images, and Figure 3.13 shows two typical examples of the original input images and the segmented images. Figure 3.10 gives the representative examples of typical failure modes. Overall, our fully automated segmentation model achieves 91% accuracy and 89% recall. The current framework still has a number of limitations that are worth discussing. First, this feature grouping approach is based on the assumption that all cell nuclei are hyper-segmented. While the superpixel approach is sufficient to capture the boundaries of most nuclei, it is not adequate for cases where two contacting nuclei have similar intensities or where there is no obvious contrast between nucleus and cytoplasm. If two nuclei are segmented into one segment in the superpixel step, the algorithm is unable to split them in the following optimisation process, resulting in under-segmentation. Secondly, the shape model in this work is defined simply as a parametric elliptical model. Although most nuclei are elliptical, highly concave or elongated nuclei can sometimes be observed in the dataset, and these usually lead to over-segmentation. Finally, the Canny edge detector may produce some internal edges in nuclei with complex textures, resulting in boundaries errors in the final segmentation.

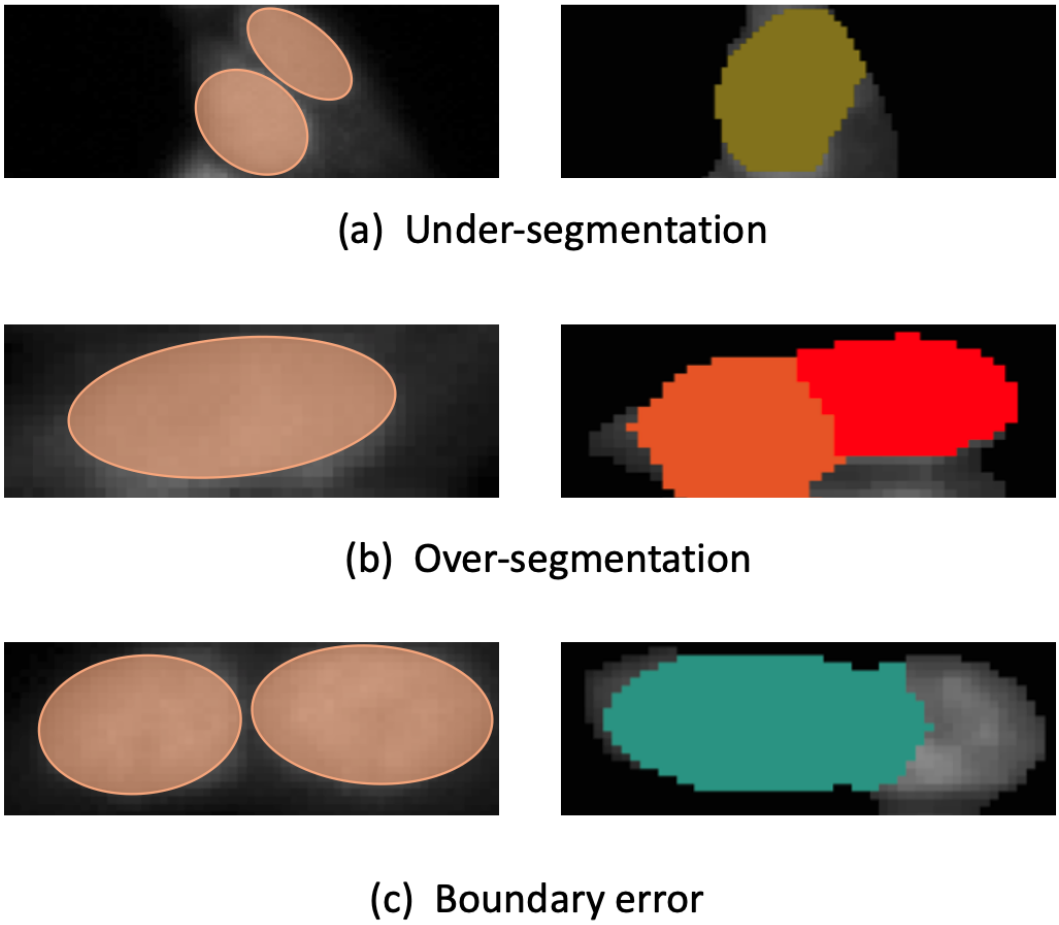


Figure 3.10: Illustration of typical failure modes in the segmentation results. In this failure analysis, there are three different types of errors. The images on the left present groundtruth labels, while the images on the right are the segmentation results generated by the proposed segmentation model. (a) Example of under-segmentation error, which describes the situations where a region is not split into the correct number of cell nuclei. (b) Example of over-segmentation error, which is defined as excessive splitting. (c) Example of boundary error. In addition to the two standard types of errors, we also looked for boundary errors, which occur when the segmentation framework does not delineate the correct boundary between a pair of touching nuclei. For instance, a detected nuclear candidate may include a portion of its neighboring nucleus, while the neighboring nucleus is not detected at all.

3.3.5 Analysis of Global Optimisation

To demonstrate the effectiveness of the proposed model-based segmentation framework, a set of visualisation experiments were conducted on the key components of the design. Starting from a high initial temperature, the model is forced to be globally optimised by gradually reducing the temperature through a linear annealing schedule. For two segments that should typically be merged together, the probabilities are uniformly distributed at the initial temperature, and become binary at the final temperature. Figure 3.11 shows the change of probabilities with respect to these two segments.

Additionally, Figure 3.12 provides an example of the results for each iteration with respect to a small patch of the image. As the results shown in Figure 3.12, most of the cell nuclei can be detected within 10 iterations. The computation time of each iteration strongly depends on the size of the input image and the number of cell nuclei in the image. In order to reduce computational costs, I directly set the final affinity score to zero if there are significant edges or long distances between two superpixel segments, without calculating other measurements in each iteration. Thus, although this is not a real-time method, the exclusion time and computational cost of the whole optimisation process is affordable and in practice can be obtained in 20 seconds to 2 minutes.

3.4 Overall Segmentation

3.4.1 Hybrid Segmentation Framework

This section discusses how the proposed model-based segmentation approach can be utilised to further improve the results obtained by convolutional networks. In our setting, very basic constraints on shape and size are being used to detect outliers, i.e., segments which result in a low ellipse fitting score or are simply too large are being processed further. As indicated in Figure 3.14, subsequently superpixel segments are generated for the corresponding local image patch. After the feature

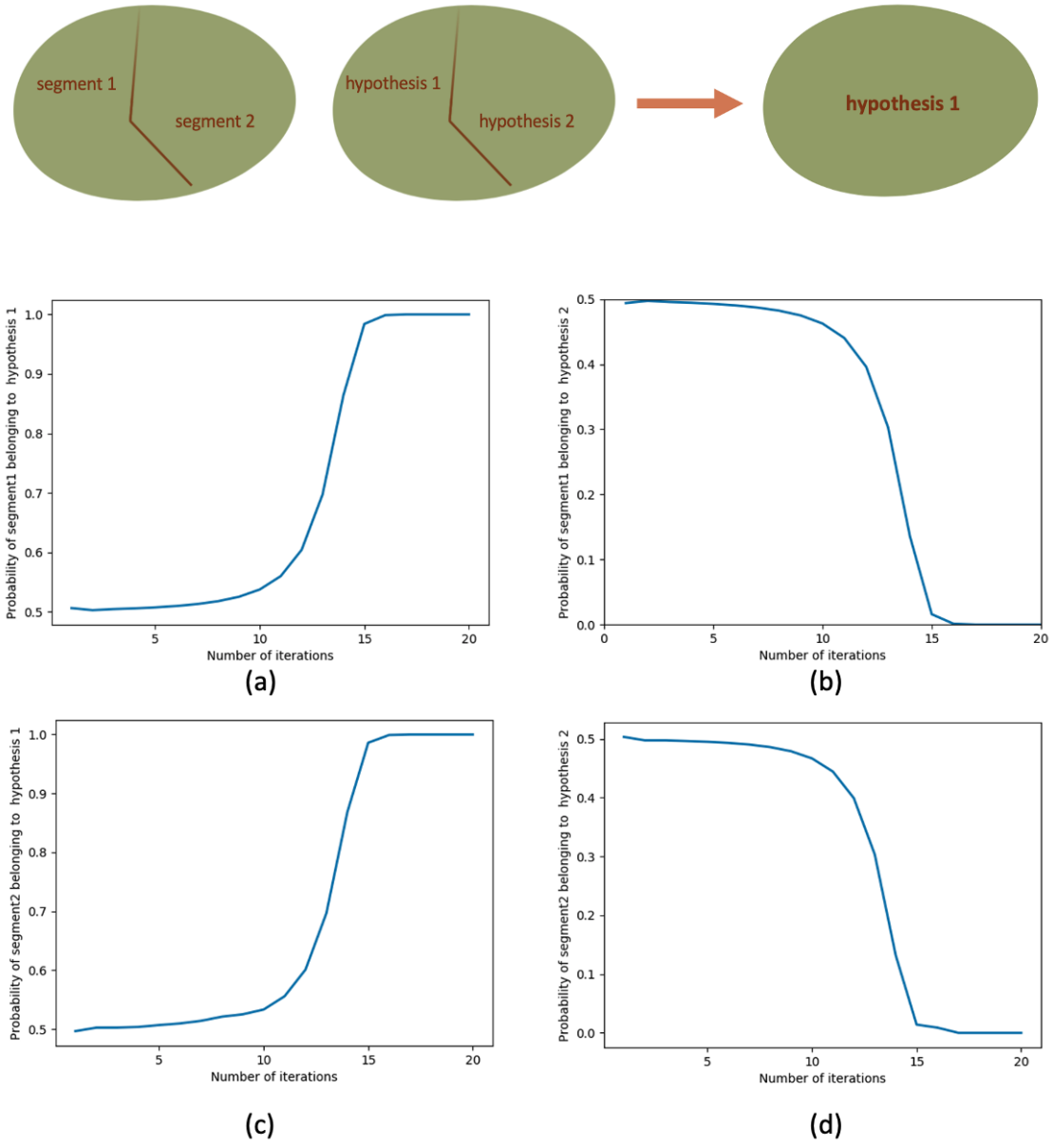


Figure 3.11: Plots of probability variance. (a) probability of segment 1 belonging to hypothesis 1, (b) probability of segment 1 belonging to hypothesis 2, (c) probability of segment 2 belonging to hypothesis 1, and (d) probability of segment 2 belonging to hypothesis 2 during the iterative process.

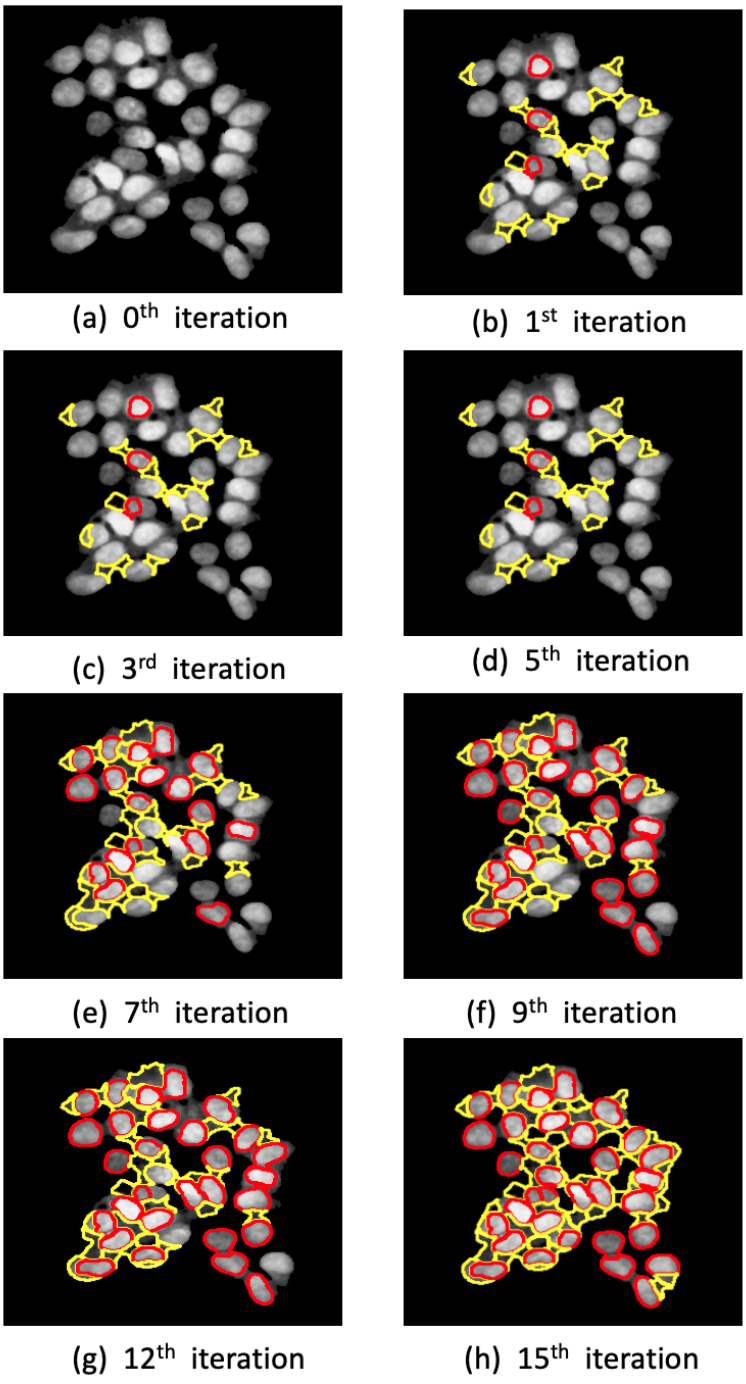


Figure 3.12: Model visualisation with respect to a small portion of the sample image. Model visualisation with respect to a small portion of the sample image. (a)-(f) describe the result of 0^{th},1^{st},3^{rd},5^{th},7^{th},9^{th},12^{th} and 15^{th} iterations respectively. The red boundaries indicate valid nuclei hypotheses, while the yellow boundaries delineate the regions of cytoplasm.

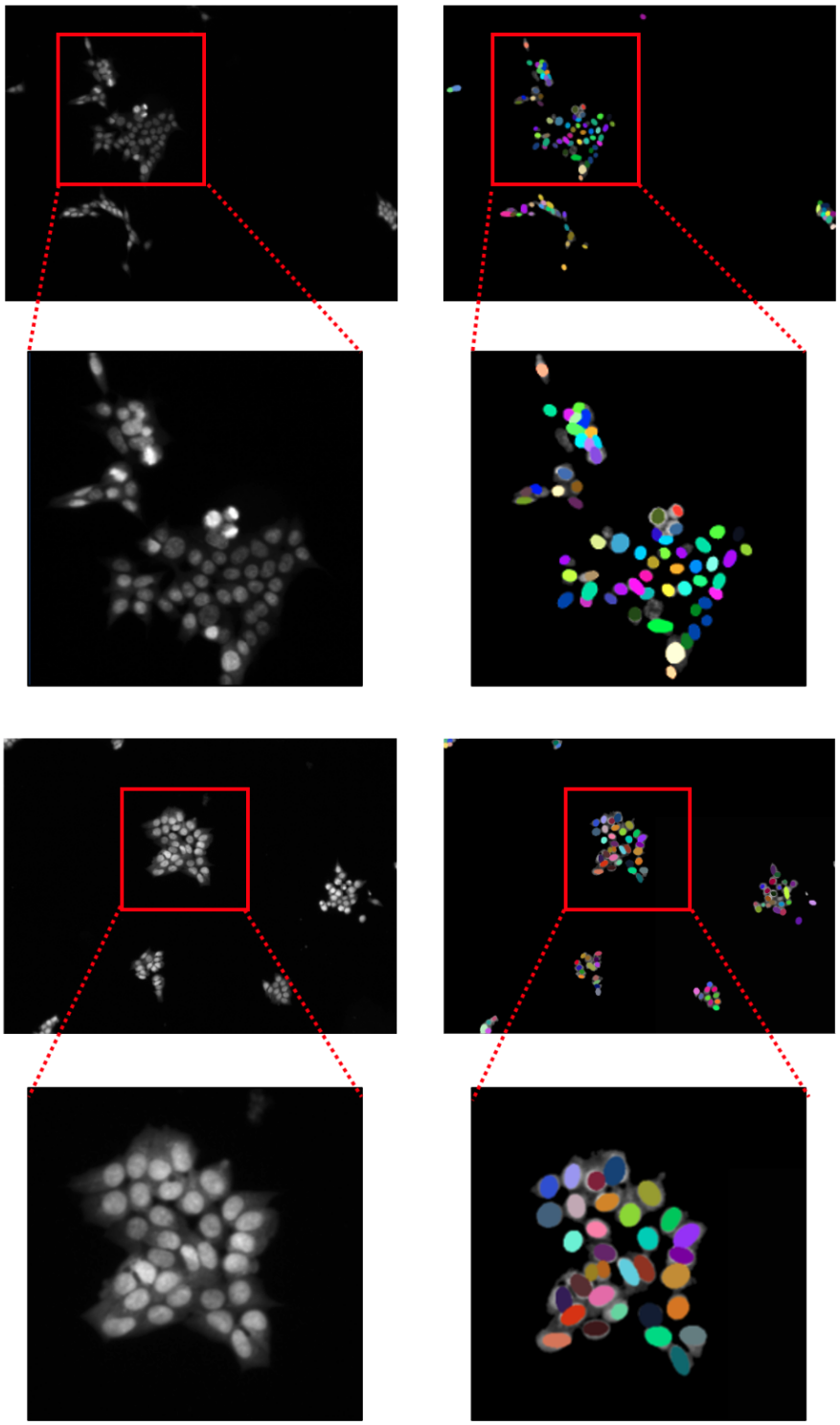


Figure 3.13: Representative segmentation results using the proposed probabilistic model. The original images and segmentation results are shown in left and right columns, respectively.

Method	Jaccard Index	Dice Index
U-Net-A	0.67	0.80
U-Net-B	0.72	0.84
U-Net-A + Probabilistic model	0.81	0.90

Table 3.4: Comparison of segmentation performances.

extraction, the global stochastic optimisation described in Section 3.3 is being applied to address the failure cases of deep networks.

Deep learning opens up new and exciting opportunities for semantic segmentation. However, in many practical applications, only a limited amount of training data is available. Here, this work demonstrates that a global probabilistic model can be used to efficiently compensate for the segmentation obtained by CNNs. I argue that the proposed framework is a very natural and efficient way of combing a data-driven and a model-based approach. The convolutional neural network robustly identifies isolated cells and adapts very efficiently to the given imaging conditions. The model-based approach, on the other hand, effectively enforces some basic shape constraints and performs global optimisation. Local appearance and spatial consistency are explicitly captured in this model with the help of the affinity functions g.

3.4.2 Experimental Results

The main results of our experiments are summarized in Table 3.4. As I employ a relatively small training set, the standard U-Net (U-Net-A) tends to give wrong predictions for densely packed nuclei. Like the training strategy used in U-Net-B, the weighted loss function can be adopted to address the problem of cell clumps in high-level vision tasks such as cell counting. However, the pixels in cell interior and cytoplasm are often predicted to be edge points, resulting in lower pixel-level segmentation accuracy. In contrast, our global model can effectively compensate for the potential segmentation failures of standard U-Net, especially when the annotated training data are limited. As shown in Table 3.4, this framework achieved a 9% improvement over existing benchmark segmentation approaches. Representative segmentation results produced by the hybrid framework are shown in Figure 3.15.

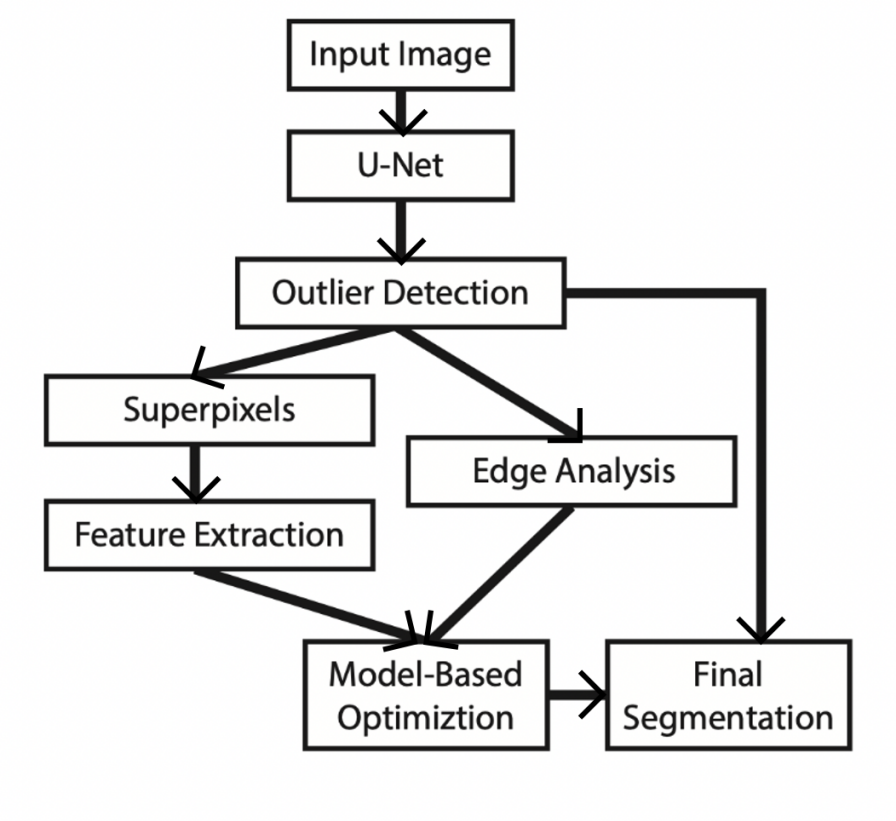


Figure 3.14: Overall segmentation algorithm. Given the segmentation results of U-Net, very basic constraints on shape and size are being used to detect the outliers (failure cases) of U-Net. Detected failure cases are handled explicitly by the proposed global probability model described in the chapter, while the successful segmentation results generated by U-Net are passed directly to the final segmentation map.

The proposed segmentation framework appears to reliably identify the densely packed cells, which effectively compensates for potential segmentation failures of standard U-Net.

3.5 Conclusions

In this chapter, the U-Net model was used to assess the performance of deep learning approaches in cell nuclei segmentation. Our experiments indicated that deep learning-based methods were not sufficient to localise the accurate boundaries for individual cells in pixel-level, especially when we only had access to a limited amount of training data. Although one could make use of additional training data

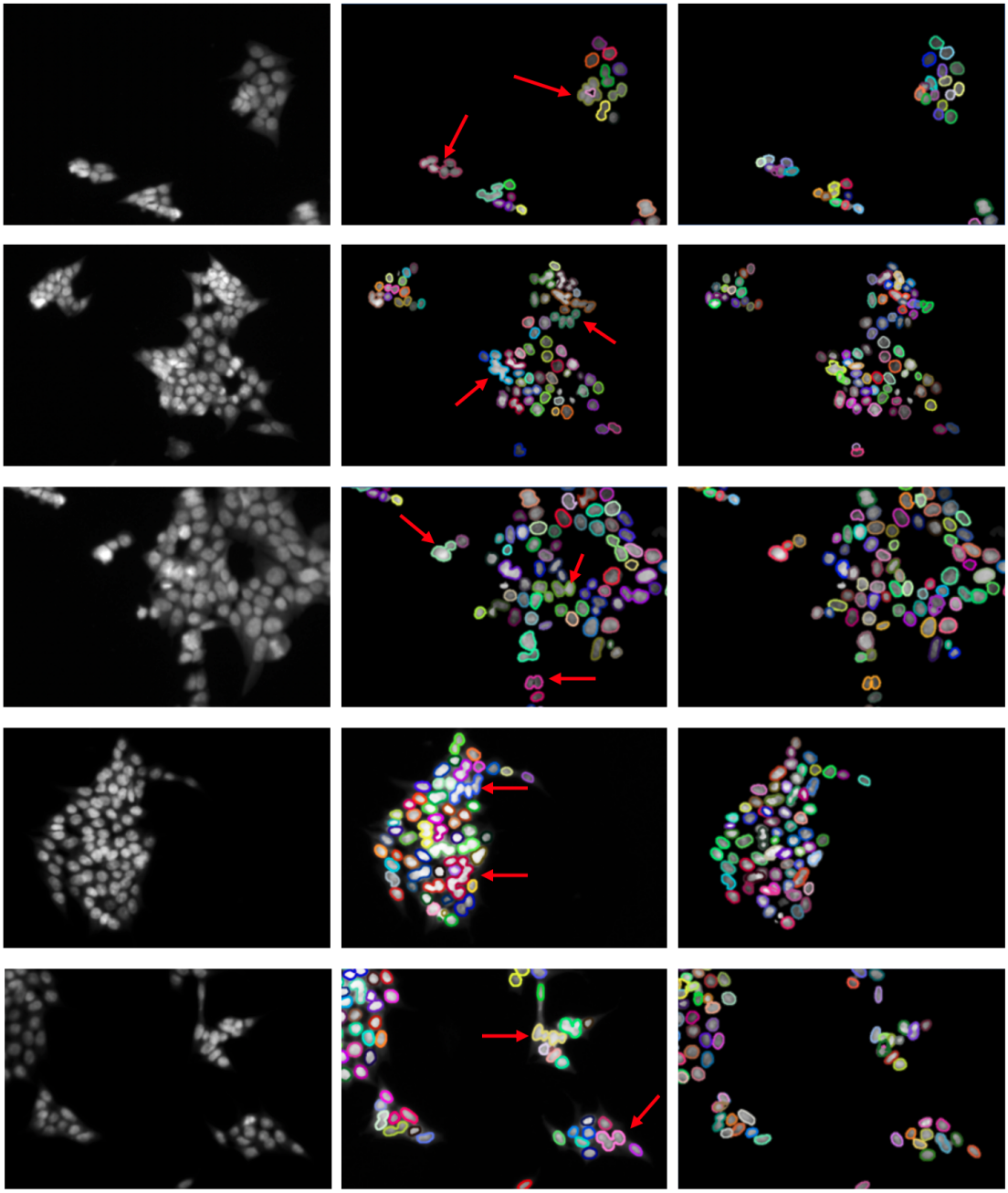


Figure 3.15: Segmentation performance of the standard U-Net and the proposed hybrid framework. The input images area shown in the first column. The second column presents the segmentation results of standard U-Net, and the third column shows the segmentation results of the proposed hybrid segmentation approach.

or data augmentation to overcome this problem, it is still important to design robust algorithms for smaller data sets.

To achieve this, this chapter proposed a new global probabilistic model for automatic segmentation of cell nuclei in fluorescent microscopy images. Specifically, the main idea of the method is to extract region-based image features from the over-segmented fragments generated by superpixel segmentation and intelligently group the features using a probabilistic model describing the anatomical appearance of the object of interest. The whole segmentation framework is fully automated and has several advantages listed below.

- **No prior knowledge is required for initialisation.** Unlike other existing model-based approaches, which initialise templates through user annotations or strong learning priors, initialisation does not require prior knowledge, user interaction, and assumed cell distributions.

- **Robust shape representation.** Although cells may have heterogeneous shapes, which usually change dynamically, the shape of most nuclei can be approximated by ellipses. This approach incorporates a parametric elliptical model that uses a fitting algorithm to ensure flexibility and robustness to different sizes, orientations, and reflections, thus addressing the segmentation of dense cell nuclei naturally.

- **Combination of various effective features.** The modelling performance of the target object is improved by simultaneously exploiting a large number of representative features. I incorporate local image cues that are heavily used in traditional data-driven approaches: raw image intensity, texture, and gradient information.

- **Global optimisation.** In order to address the limitations of traditional merging methods, the corresponding energy minimisation was globally performed on the entire image.

- **Generalisation.** The proposed model is not limited to one particular cell imaging data set. It can be easily extended to other types of biomedical images where the high-level prior knowledge of objects is available.

The global probabilistic model was applied to fluorescence microscopy images, which demonstrates that the model is capable of accurately segmenting densely packed populations of cells, while the time and computational cost are both affordable in practice. Our experiments also showed that the combination of U-Net and the global optimal model leads to significant improvements in segmentation accuracy. However, the whole segmentation framework can not be trained in an end-to-end fashion. The computational complexity and segmentation accuracy also heavily rely on the results of the superpixel algorithm. To address the limitations, subsequent chapters will focus on developing end-to-end networks to learn a more informative feature representation for small datasets.

4

Fine-grained Fungi Species Classification through Deep Attention

Contents

4.1 Introduction 85
4.2 Method 90
4.2.1 Attention Mechanism 90
4.2.2 Feature fusion 93
4.2.3 Multi-task Loss Function 97
4.3 New Clinical Fungi Dataset 98
4.3.1 Clinical background 99
4.3.2 Morphological Features 100
4.3.3 Data Acquisition 103
4.3.4 Dataset Description 104
4.4 Experiments and Results 105
4.4.1 Implementation details 105
4.4.2 Quantitative Comparison with Competing Methods . . 106
4.4.3 Comparative Evaluation of Attention Mechanisms . . . 109
4.4.4 Visualisation of the Attention Map 110
4.5 Conclusion 113

This chapter will presented the first step towards recognising fine-grained structures in biomedical images. Deep learning-based vision systems mostly deal with a high number of low-resolution images, whereas fine-grained recognition of biomedical images requires higher resolution information. To bridge this gap, this

chapter describes a simple yet effective attention-based multi-task learning scheme which simultaneously integrates local high-resolution and global structural features in an end-to-end fashion. In addition to the development of new algorithms, this work also constructed a new clinical mycology dataset of fungi images comprising of five species that most commonly affect humans.

Originality: This chapter presents a new clinical fungi dataset. I independently came to Peking Union Hospital and collated the whole-slide images. The most informative regions and the corresponding image labels were provided by collaborator Wenhang Yang and Yan Xu. I independently proposed the design of this attention-based fine-grained classification pipeline. I independently designed and implemented the attention module, hard sampling and feature fusion scheme in Python and MXNet. The code for the soft sampling study was built on top of [119]. I also independently performed validation experiments and related visualisation work. The main ideas were also presented in:

Mengran Fan, Tapabrata Chakraborti, Eric I-Chao Chang, Yan Xu, Jens Rittscher. Fine-grained multi-instance classification in microscopy through deep attention. *Proceedings of the International Symposium on Biomedical Imaging (ISBI)*, 2020. - Oral presentation.

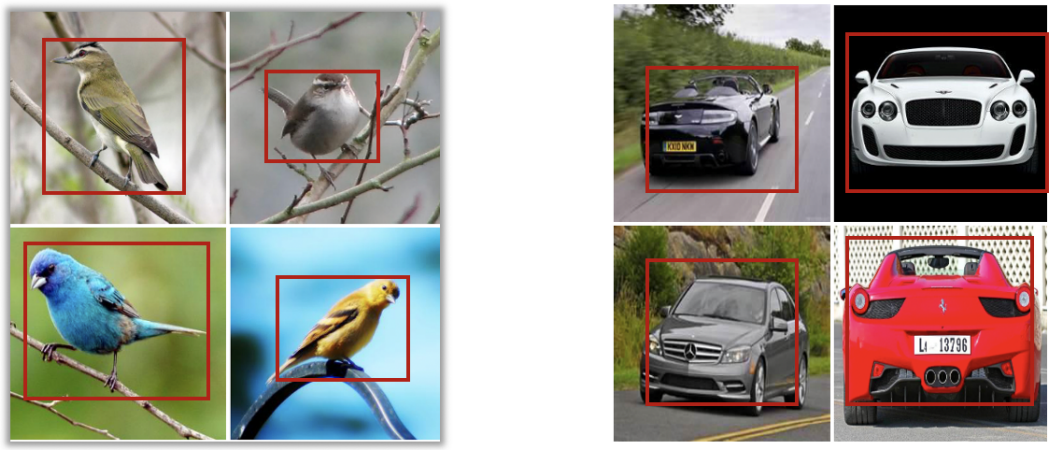
4.1 Introduction

In the past few years, deep convolutional networks have proven to be superior in classifying basic image classes on large datasets. Unlike general image recognition, many real-life applications may involve classifying subclasses belonging to the same basic class with limited data, such as species identification in biomedical image analysis. This is still an open issue in computer vision and biomedical image analysis, as fine-grained image recognition should be able to localise and identify subtle visual distinctions in critical areas within subordinate categories, especially when objects are not properly aligned, such as variations in background and pose. A large number of early works addressed this problem by explicitly using manually labelled bounding boxes or even part annotations to describe part-based feature

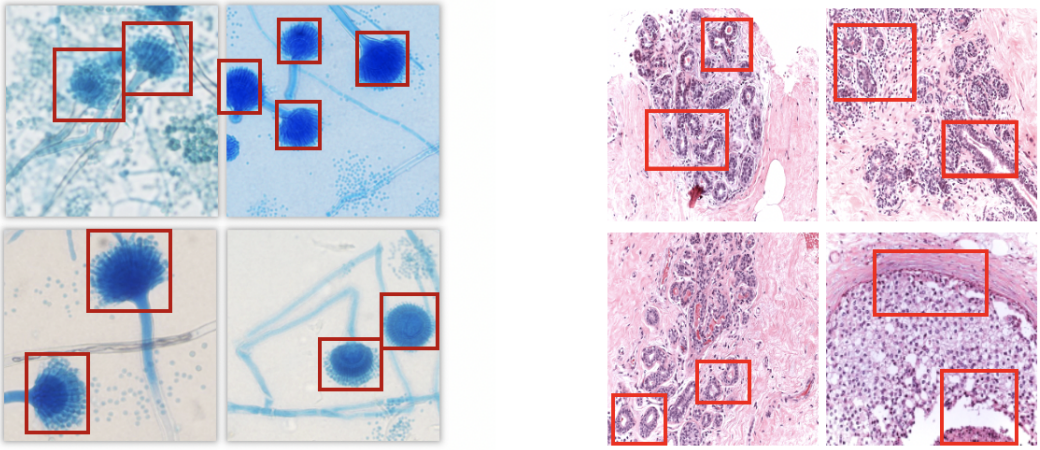
representations in both training and testing phases [137–139]. While great progress has been reported, such annotations are demanding, leading some works to consider utilising local annotations only during the training phase, rather than the testing phase.[144, 145]. For most fine-grained recognition tasks, however, even this setup is not ideal, as the local annotations needed during the training process are still heavily dependent on human intervention, which is expensive and subjective. Recently, there has been emerging work aimed at solving the problem of fine-grained image recognition with only image-level labels, and significant progress has been made on large-scale natural images [112, 119, 153].

Despite the important advances in computer vision, it is usually challenging to achieve the same success on specific biomedical image classification tasks as it poses certain interesting challenges [75, 160]. Firstly, due to the cost of data acquisition and the limited availability of specimens, well-organised datasets in the biomedical domain usually tend to be small, which directly limits the representation ability of deep networks. The main reason is that the current state-of-the-art convolutional neural networks (CNNs) are capable of extracting semantically meaningful features and complex input patterns on large-scale datasets. Despite their success, it is shown that when training data is limited deep networks may be prone to overfitting to training set biases (e.g. the changes in viewpoint, background and illumination) while ignoring the fine-grained local details, eventually resulting in poor generalisation performance. To address this problem, this chapter introduces a simple and lightweight attention module to reject the useless deep feature maps that highlight the background clutter or other irrelevant regions. With more discriminative feature maps, I can then force the network to make the predictions based on biologically or medically interpretable features.

Second, especially when working on the microscopic scale, multiple instances (e.g. glands, vessels, or crypts) need to be considered. This seriously restricts the adaptation of existing methods in the fine-grained classification of natural images. For instance, some samples (Figure 4.1 (a)) were randomly selected from the most popular fine-grained datasets in computer vision (CUB-200-2001 [161] and Stanford



(a) natural image



(b) biomedical image

Figure 4.1: Classification challenges in biomedical imaging. Compared to large-scale datasets on natural images (a), well-organised biomedical datasets (b) tend to be small and require expensive human expert annotations. Secondly, while we typically expect one centre-aligned instance (red box) in natural images, there are often multiple discriminative instances (red boxes) in biomedical images, which poses another challenge for feature learning. Finally, high-resolution information and interpretability need to be considered when developing a reliable medical image analysis system.

Cars [162]), where there is mostly one centre-aligned instance in an image. Although a large number of strategies have been proposed to detect the discriminative parts (e.g., head, belly for birds) in such images, the size and layout of the detected components are almost identical for each image. In comparison to natural images, biomedical images (Figure 4.1 (b)) may have a wide variety of discriminative instances (regions) with different sizes and densities, leading to more complicated structural information and a larger within-class variation. This motivates the need to investigate methods for building comprehensive and discriminative feature representations, which can not only capture global structural information, but can also extract local features from multiple discriminative instances simultaneously. Regarding this issue, the chapter proposes a multi-instance fine-grained recognition framework that is capable of simultaneously detecting multiple attention regions with strong discrimination capabilities.

Finally, biomedical images are usually characterised by high-resolution information, where most deep learning-based methods directly perform the down-sampling or follow a patch-based approach due to memory constraints and high computational costs. However, the downsampled images run the risk of losing discriminatory features from the high-resolution version, e.g., nuclei or lymphocyte information. Alternatively, some methods [73, 74, 96] first divide high-resolution images into locally equal-sized patches and train a two-stage framework for the patch- and image-level classification. Although patched-based models can be trained on a large number of local high-resolution patches, the deep networks tend to ignore the contextual information, i.e., the spatial relationships between neighbouring patches. Due to the lack of global contextual information, the inference is irrelevant to the underlying instance organisation, which eventually degrade the final classification performance. To this end, the proposed model first downsamples the original images and then feeds the lower-resolution images to a context-aware network for capturing global structural features such as object organisation or the number of objects. After exploiting the trained network's attention, the most discriminative regions can be localised without extra labelling effort. As shown in Figure 4.2,

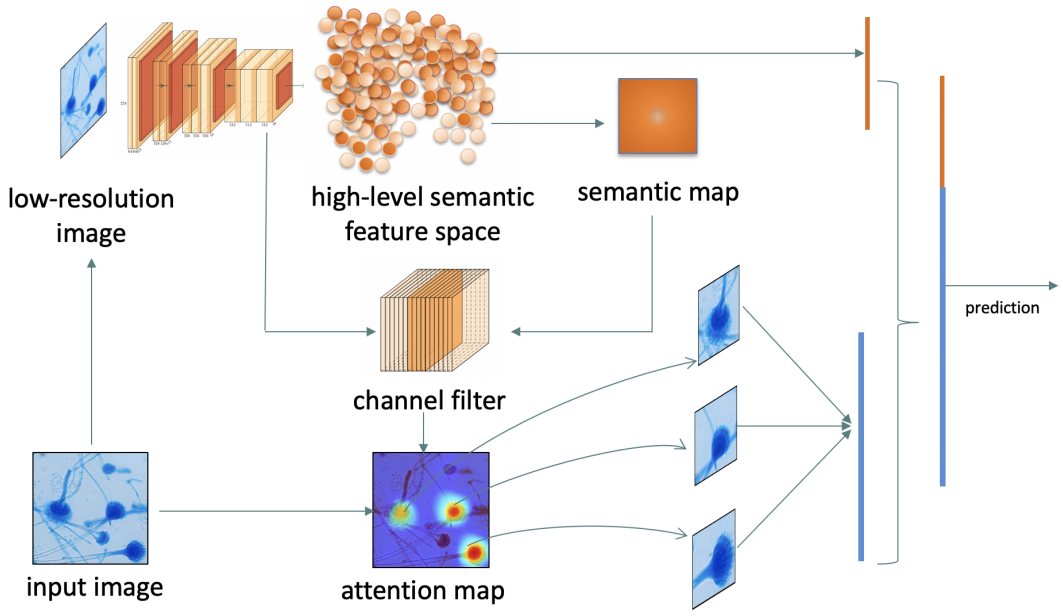


Figure 4.2: An overview of the proposed fine-grained classification framework. Using only image-level labels, the proposed model learns to localise multiple discriminative instance simultaneously and to integrate the global structural information with local high-resolution features for fine-grained classification.

the corresponding patches were directly cropped from the original image to take advantage of the relevant high-resolution information. In contrast to conventional patch-based model, the proposed attention-based framework is not subject to pre-defined patches (i.e., the size of the patches, the sampling scale), and provides a more complete region for the instance of interest.

Figure 4.2 presents an overview of the proposed fine-grained classification framework (details of the method are explained in Section.4.2). Validation experiments are carried out on a new fine-grained multi-instance biomedical image classification task: fungi species recognition. Our method outperformed the state-of-the-art weakly-supervised fine-grained classification approaches and existing attention-based methods, demonstrating the effectiveness of the proposed framework. The main contributions of this chapter are:

- A lightweight attention module where multiple discriminative instances can be localised simultaneously at high resolution without requiring any part annotations or redundant region proposals.

- A robust but lightweight end-to-end trainable deep network for biomedical image classification where the global structural features and local instance features are seamlessly fused.
- A multi-task learning scheme that dynamically controls the weights of member modules and enforces the network to learn consistent features.

The outline of this chapter is as follows. The attention-based multi-instance localisation and the feature fusion module are described in Section.4.2.1 and Section.4.2.2, respectively. Section.4.2.3 defines a multi-task loss function that enables the model to jointly learn multi-instance localisation and image classification. The clinical motivation and details of the new fungi dataset are introduced in Section.4.3. Validation experiments with implementation details are reported and discussed in Section.4.4. It concludes, in Section.4.5, with a discussion of the limitation of the work and highlight areas of possible future work.

4.2 Method

Here, the section will give an overview of the proposed framework, followed by a detailed description of each stage. Aiming to incorporate global context information and local high-resolution features, I propose a novel fine-grained multi-instance classification scheme (Figure 4.3) that consists of two main modules: (i) a novel attention mechanism for discriminative instances localisation; and (ii) a feature fusion scheme that consolidates the global and local features to facilitate the final predictions.

4.2.1 Attention Mechanism

The attention mechanism has been extensively studied in the fields of computer vision and natural language processing, such as image inpainting [108], image and video captioning [111], semantic segmentation [110] and machine translation [109]. Recently, various methods incorporate attention mechanism to enhance the feature representation ability of CNNs, which demonstrated that the attention mechanism is

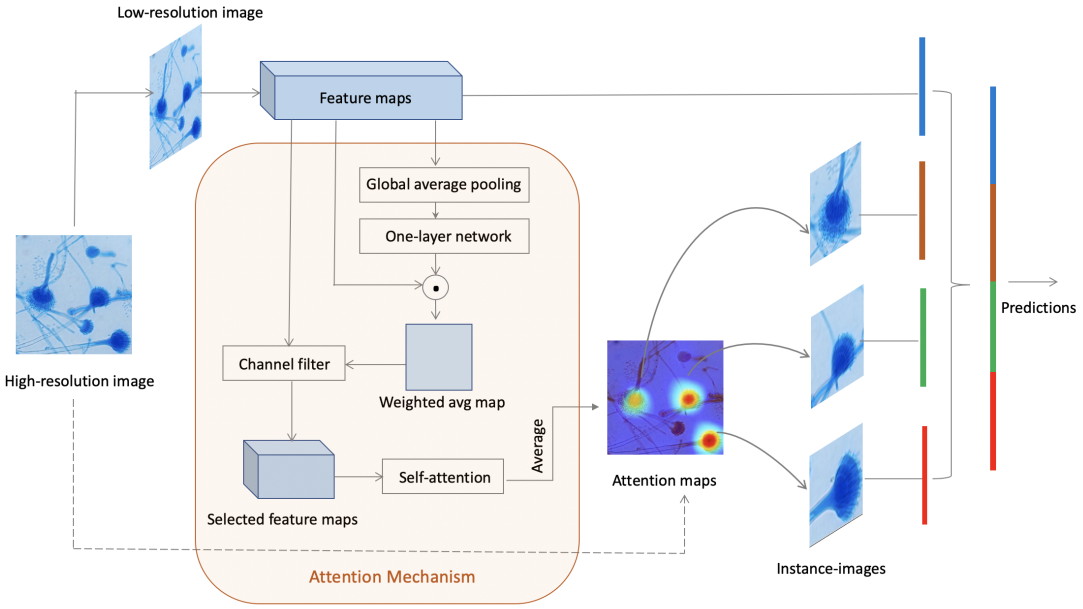


Figure 4.3: Framework for the proposed multiple instance fine-grained classification pipeline. The network consists of two main modules: (i) a novel attention mechanism for discriminative instances localisation; and (ii) a feature fusion scheme that consolidates the global and local features to facilitate the final predictions.

effective in highlighting the salient regions and discarding the irrelevant information. The proposed attention mechanism aims at automatically localising discriminative instances in fine-grained images without extra supervision or redundant region proposals. As shown in Figure 4.3, the whole model can be divided into three procedures: (a) semantic modelling framework, which groups the channels via weighted averages with learnable weights; (b) channel filter for rejecting the feature maps highlighting the irrelevant regions; (c) lightweight self-attention model, which globally optimises the selected feature maps.

Semantic Modelling Framework: It is aimed at generating an aggregated global feature map with strong semantic meaning (roughly highlighting the main objects but discarding the noisy background). It was built by calculating a weighted average of original feature maps obtained from the low-resolution images. This approach rests on the assumption that if many channels focus on the same region, we could expect this region to be part of objects.

To learn the importance weight for each channel, channel descriptors are

constructed by global average pooling. Then importance weights were fed to a one-layer network. The global semantic map is constructed by computing a weighted average using the learnt weights. With features maps $\mathbf{X} \in \mathcal{R}^{H \times W \times C}$ the aggregated semantic map can be written as:

$$w = \sum_{p=1}^{C} F(x_p, \delta(\frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_p(i,j))), \tag{4.1}$$

where $\delta(.)$ denotes the feature transform with fully-connected layers and $F(.,.)$ refers to a rescaling function to fuse the importance weights and original feature maps.

Channel Filter: In terms of the large-scale image dataset, background or irrelevant parts can be naturally discarded in a deep neural network. However, for the limited microscopy imaging data, an extra operation selecting effective deep descriptors would be necessary for noisy information removal and improved object localisation. Here, this section introduces a simple yet effective method to filter out the noisy feature maps. With the global semantic map w that roughly indicates the position of main objects, the pairwise similarities between itself and each original feature map x_p can be calculated based on their activation responses. Consequently, these similarity values can be sorted in the descending order and only the first $N(N < C)$ channels can be selected for subsequent transformations. After conducting experiments with different values of N, it can be found that the classification accuracy is robust to values of N greater than 100. Therefore, I chose only the first 100 channels in this step to reduce the computational cost.

Self-attention Module: After filtering, the selected low-dimensional feature descriptors are obtained. As mentioned before, it is important to attend all relevant parts of the object to keep sufficient fine-grained information. Therefore, the model further recalibrates the activation responses by employing a self-attention module [119], where global long-range channel interactions can be captured explicitly to enhance the object boundaries and highlight the meaningful regions. As shown in Figure 4.4, the self-attention module can be formulated as:

$$\mathcal{M} := (XX^T)X, \tag{4.2}$$

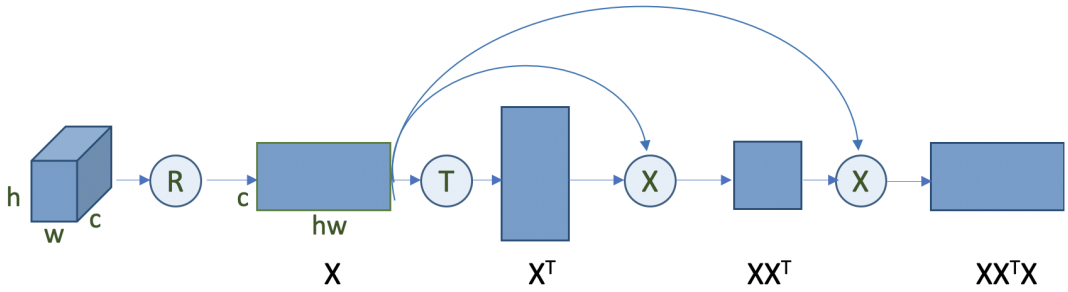


Figure 4.4: Self-attention module. Given a feature vector $\mathbf{X} \in \mathcal{R}^{h \times w \times c}$, where h, w, c stand for the height, width and channels respectively, the self-attention module XX^TX is used to attend to the entire region of the biomedical object.

where XX^T refers to the pairwise inter-channel similarities, which indicates the spatial relationship among channels. To enhance the robustness and consistency of the attention map, the normalised trilinear attention further integrates the original convolutional feature maps, conducting dot production over XX^T and X.

To further improve the effectiveness, spatial softmax normalisation is adopted as:

$$\mathcal{M} := \mathcal{N}(\mathcal{N}(X)X^T)X, \tag{4.3}$$

where $\mathcal{N}(.)$ indicates softmax normalisation. So $\mathcal{N}(X)$ and $\mathcal{N}(\mathcal{N}(X)X^T)$ are spatial normalisation and channel-wise relationship normalisation, respectively. Finally, all generated attention maps are added along the depth direction for the aggregated global attention map, which can be used to localise the main objects. The computational burden of this self-module is very low as it is only applied to the selected low-dimensional feature maps.

4.2.2 Feature fusion

When multiple informative regions are proposed, there are three different feature learning approaches for fine-grained classification: (i) soft sampling where regions with high responses are sampled more densely, (ii) hard sampling that crops the patches from the original image and (iii) deep descriptors selection, which directly selects the useful deep descriptors in high-level feature maps. Compared to deep descriptors selection, resampling from original images can make good use of the

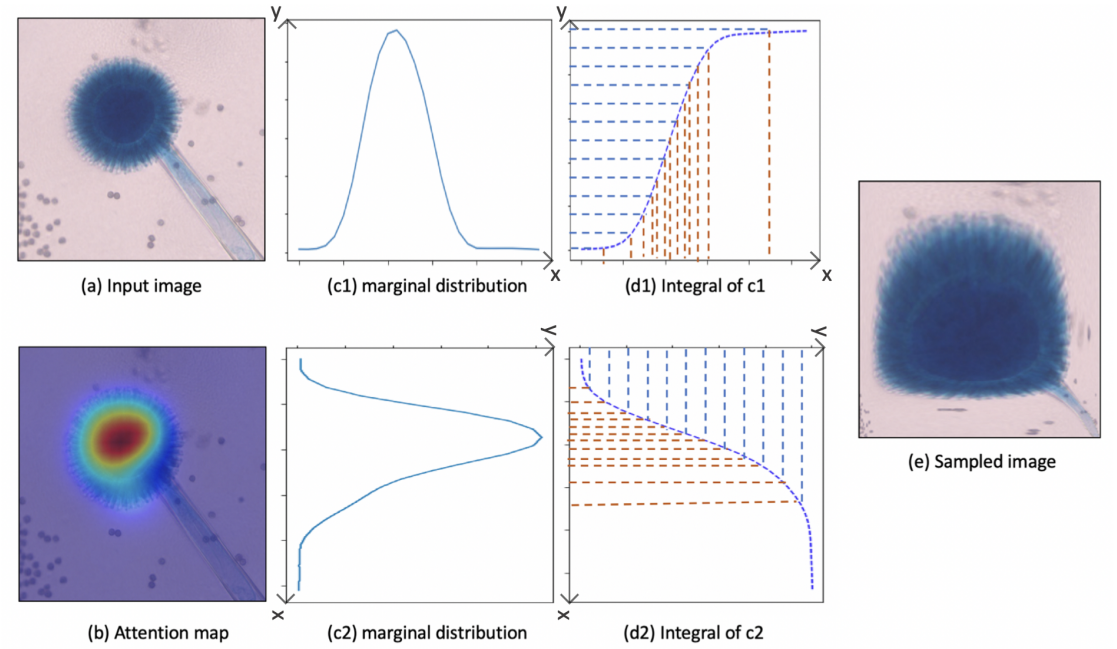


Figure 4.5: The mechanism of the attention-based soft sampling. (a) is the original full resolution image and (b) is the attention map generated by the proposed attention module. (c1) and (c2) are the marginal distributions over x and y axis, respectively. (d1) and (d2) are the cumulative distribution functions of (c1) and (c2). (e) illustrates the final sampled images.

high-resolution information, and directly enhance the reliability and interpretability by providing extracted patches to a human expert as well as final predictions. Therefore, in this work, I focus on sampling strategies ((i) and (ii)), which aim to sample the informative pixels from the original high-resolution image.

Soft sampling strategy: The effect of the sampler is to discover and zoom in on relevant regions from the underlying high-resolution data in a way that preserves the information relevant to the task at hand. Compared to hard sampling, which may crop multiple patches, the basic idea of soft sampling is to produce only a single sampled image, where regions with large attention values are sampled more densely.

Specifically, the sampler g takes as input the attention map S along with the original high-resolution image I and computes J = g(I, S). Inspired by the inverse-transform sampling technique[163], this sampler considers the attention map as a probability mass function and uses it to calculate the mapping between the coordinates of the sampled image and the original image. This sampling process can

be divided into two stages. In the first stage, the sampler decomposes the generated attention map into two dimensions and calculate the cumulative distribution function (CDF) over x and y axes to avoid spatial distortion. In the second stage, the inverse function of the distribution function was calculated to obtain the sampled points.

Such a sampling mechanism is illustrated in Figure 4.5. Given an attention map S, I first normalise the attention weights and decompose the map into two dimensions by calculating the max values over x axis (c1) and y axis (c2). As shown in (d1) and (d2), the cumulative distribution functions can be expressed as the integral of the probability mass functions (c1) and (c2). Note that for discrete distributions, calculating the cumulative distribution functions is not as difficult as for continuous distributions: we can simply sum the probabilities for each point in the distribution.

$$\begin{aligned} \mathcal{F}_x(n) &:= \sum_{j=1}^{n} \max_{1 \le i \le w} S_{i,j}, \\ \mathcal{F}_y(n) &:= \sum_{i=1}^{n} \max_{1 \le j \le h} S_{i,j}, \end{aligned} \tag{4.4}$$

where w and h are the width and height of the attention map, respectively. Finally, the inverse functions of (d1) and (d2) were calculated to generate the sampled image J from the full resolution image I:

$$J(I, S)_{i,j} := I_{\mathcal{F}_x^{-1}(i), \mathcal{F}_y^{-1}(j)} \tag{4.5}$$

In the sampling process, this method uniformly samples points over the y axis (blue arrows shown in (d1) and (d2)), and obtain the corresponding values over x axis (red arrows shown in (d1) and (d2)). The final sampled image is shown in (e). It can be observed that the regions with large attention values are allocated with more sampling points and thus sampled densely. This sampling strategy is capable of recovering small image features or subtle differences between related species with a limited computational cost. More representative examples of this soft sampling strategy are shown in Figure 4.6.

Hard sampling strategy: However, it can be found that the concept of soft sampling strategies is not well suited to biomedical images, such as those

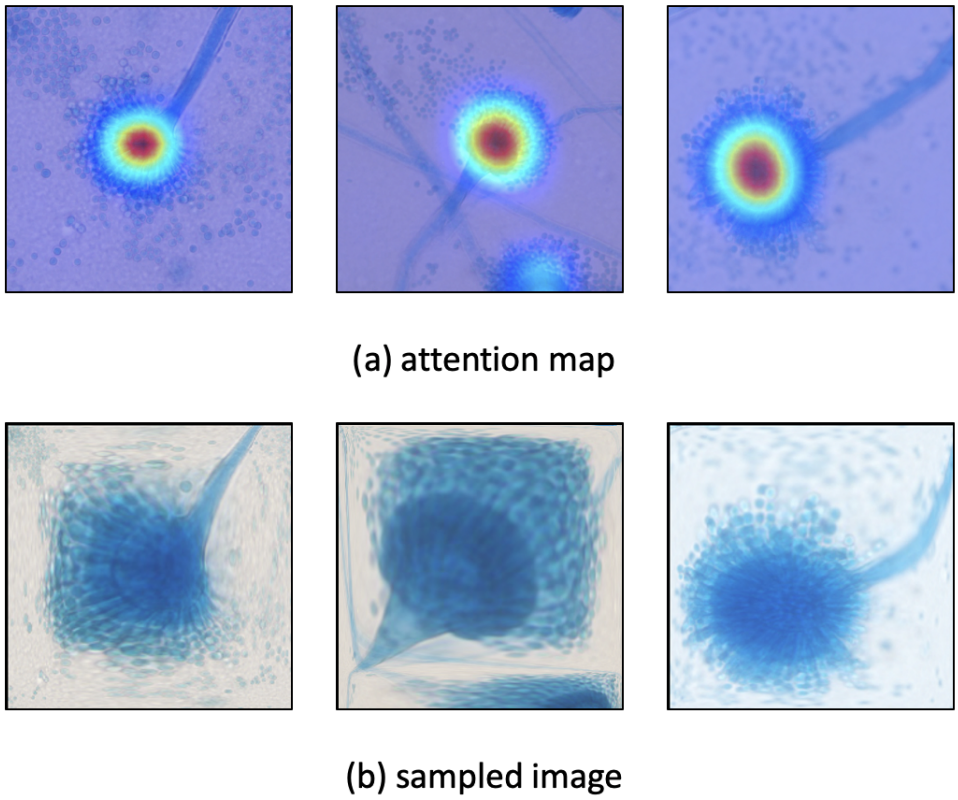


Figure 4.6: Examples of the attention-based soft sampling.

in our application, in which there may be multiple discriminative instances. As shown in the failed example in Figure 4.7, most soft sampling methods result in spatial distortions and therefore fail to preserve the shape information of the instances. To that end, the method directly adopts the conventional hard sampling strategy to generate multiple informative instance-level patches. While uniform hard sampling is simple and effective in this task, it can preserve the necessary shape information and discard other irrelevant regions. In particular, thresholding and connected component analysis were performed on the global attention map. With the bounding box positions and corresponding sizes, N sampled images containing the main objects were generated by cropping squares from the original high-resolution images. The cropped patches were scaled up to a larger resolution (e.g., 336×336) and fed to a shared CNN-based network for the feature extraction of local details. Finally, these instance-level features and high-level structural

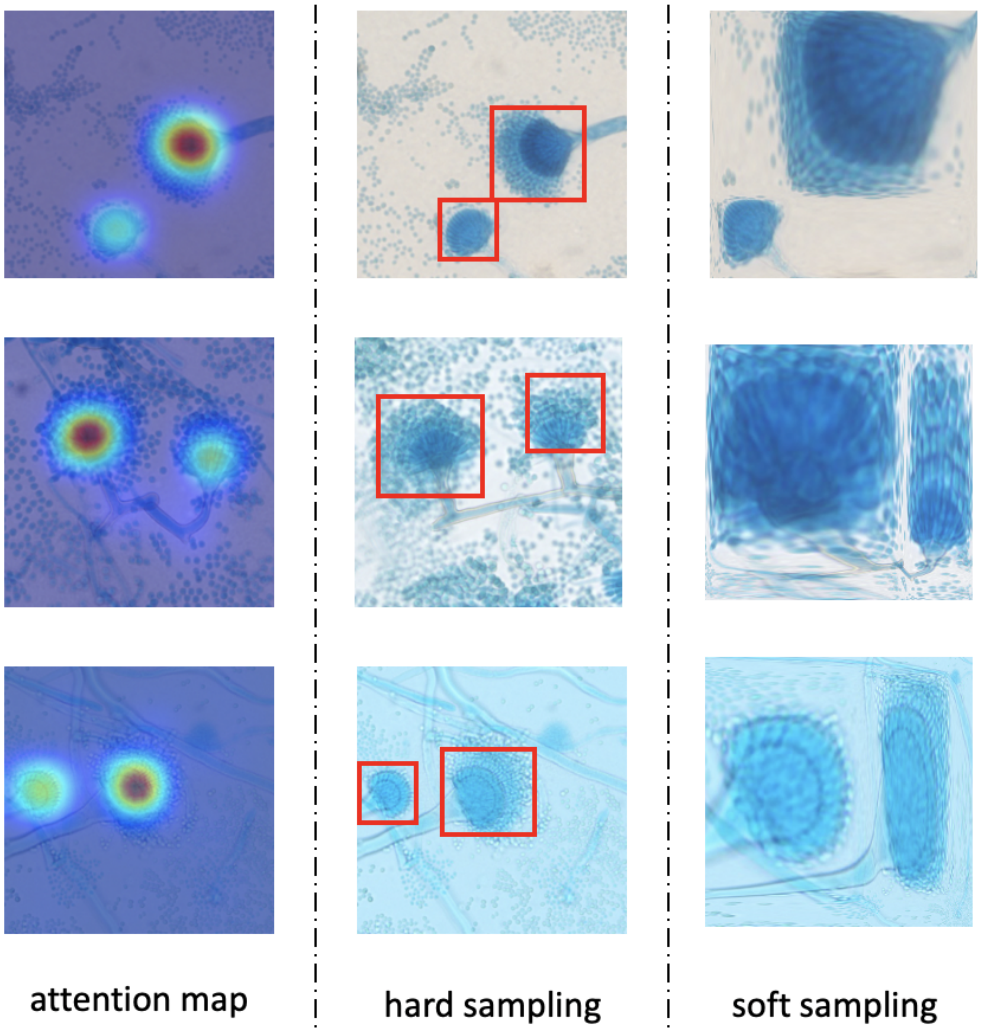


Figure 4.7: Failure cases of soft sampling. It can be seen that soft sampling may result in spatial distortion when it applied to multi-instance cases. To keep the shape information of multiple instances, hard sampling were conducted to generate informative regions.

features were consolidated for final predictions.

4.2.3 Multi-task Loss Function

Different from standard two-stage frameworks, which consist of two separate networks, the proposed multi-task framework aims to enable the model to jointly learn discriminative patch-level features and global image-label contexts in an end-to-end fashion. Our multi-task model is divided into two separate task: global

context-aware feature learning and multi-instance-based image classification. This work defines loss functions for the global context-aware module and the image classification module based on the standard categorical cross-entropy between the output of the softmax layer,P_g, and the ground truth label of the input image, $Y^\star$:

$$\begin{aligned} L_{global}(P_g, Y^\star) &= -\sum_{i=1}^{C} p_g \log Y^\star \\ L_{fusion}(P_f, Y^\star) &= -\sum_{i=1}^{C} p_f \log Y^\star \end{aligned} \tag{4.6}$$

where C is the number of categorical labels in the dataset (C=5, in the fungi dataset). Finally,, the proposed multi-task loss $\mathcal{L}$ is denoted as:

$$\mathcal{L} = \lambda L_{global}(P_g, Y^\star) + (1 - \lambda)(L_{fusion}(P_f, Y^\star)) \tag{4.7}$$

where λ is a weight parameter that is initialised as 1 and gradually decreased during the training process. As a result, the loss of the global context-aware module converges faster than the loss of the multi-instance classification module, which enables the network to focus on extracting global high-level semantic features at the outset, gradually increasing the contribution of discriminative instance-level features during the training process.

4.3 New Clinical Fungi Dataset

As part of this work, I constructed a new clinical fungi dataset comprising of five species that most frequently affect human health as discussed in the next section. To the best of our knowledge, this project is the first attempt for bringing deep learning-based approaches to fungi species recognition. This section will begin by outlining the clinical background of fungi species recognition and the motivation for incorporating deep learning-based methods. The acquisition procedures for the data and a discussion of the details are described in Section.4.3.3 and 4.3.4, respectively.

4.3.1 Clinical background

In recent years, the frequency and severity of fungal infections have increased tremendously. Globally, more than 300 million people suffer from serious fungal infections, with 25 million of them at high risk of death. In particular, Aspergillus species are the major invasive fungal pathogens that cause a group of diseases called aspergilloses, such as pulmonary aspergillosis and endocarditis. These infections usually affect the respiratory system, but the symptoms and severity vary widely. Although rare in healthy people, the most severe form of aspergillosis, known as invasive aspergillosis, is fatal and only occurs in people with lung disease or a weakened immune system due to cancer chemotherapy, bone marrow transplantation or immune system disease.

Depending on the type and kind of pathogens, treatment may involve observation, antifungal medications or, in rare cases, surgery. Generally, early diagnosis and accurate identification of these species in clinical laboratories are vitally important for effective treatment due to the high frequency of antifungal drug-resistant isolates of such Aspergillus species. However, there are many barriers to clinical diagnoses, such as the increasing scarcity of clinical mycologists, high costs, and poor sensitivity and specificity.

In most clinical laboratories, the routine identification of Aspergillus species is traditionally based on methods such as direct microscopic examination of clinical samples, histopathology, and culture. Such approaches feature by clinical mycologists with a high level of specific mycological expertise. However, due to AIDS (acquired immune deficiency syndrome) and the use of highly immunosuppressive treatments for various diseases, these fungal infections occur most often in the geographical regions with limited senior mycologists and related medical resources. As a consequence, such limitations have led to an alarming rise in the number of undetected infections and delayed treatment in these regions.

On the other end, the number of fungi that clinical mycologists must identify continues to grow in recent years, which is likely to be many times the number described in textbooks. Therefore, the growth in awareness of medically relevant

fungi and the dearth in the number of highly skilled mycologists have inspired researchers to develop innovative technologies that go beyond traditional approaches and meet the needs of resource-limited clinical environments. In particular, emerging technologies such as deep learning-based image analysis methods may offer new options for automatic and highly accurate identification of current fungal species and provide more insight into the antifungal therapy. The purpose of the study was to introduce deep learning-based approaches to Aspergillus species recognition and investigate the strengths and weaknesses of state-of-the-art fine-grained classification methods when applied to this specific task.

4.3.2 Morphological Features

As discussed in the previous section, accurate identification and characterisation of Aspergillus species may be important due to the differing susceptibilities of different species to multiple antifungal drugs. Thus, knowledge of the individual species may assist the clinician in investigating an outbreak and selecting the most appropriate antifungal treatment. Although the investigations into new diagnostic technologies have been carried out extensively, clinical microbiology laboratories still rely heavily on morphology-based identification methods to accurately identify clinically relevant Aspergillus species to species level. In order to develop reliable diagnostic techniques, it is essential to clarify these phenotypical characteristics associated with species-level recognition of clinically relevant Aspergillus species. Clinically, several Aspergillus species are frequently involved in human diseases such as A. flavus, A. fumigatus, A. nidulans, A. terreus and A. niger. The principal morphological characteristics of individual species were outlined in [164, 165]. Results of these studies showed that a convenient universal criterion for fungal species identification is a range of morphological features associated with conidial heads which consist of the vesicles, metulae (if present), phialides and conidia. Important diagnostic criteria include identification of their structures and characteristics such as size, colour, shape and/or mode of attachment. Figure 4.8 presents a schematic diagram of the structure of Aspergillus species and key morphologic features include:

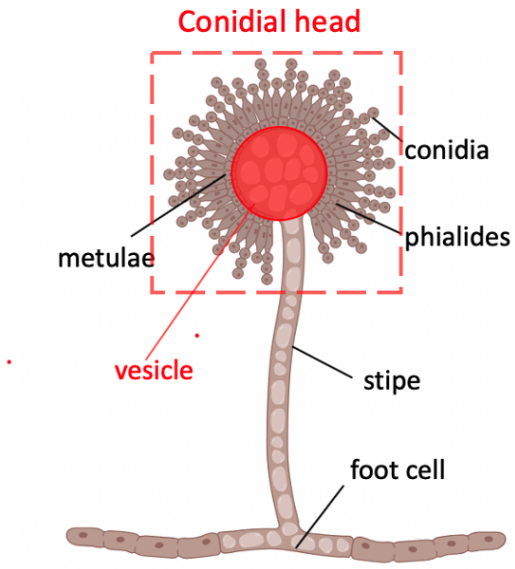


Figure 4.8: A schematic diagram of the structure of Aspergillus species. A more detailed description of the structure is provided in the text. (The figure was created with `https://BioRender.com`.)

- **size and shape of vesicles:** the stipe terminates in a swollen cell called a vesicle, which can be globose, subglobose, dome-shaped or hemispherical. Differences in vesicle appearances are vitally important for differentiating Aspergillus species;

- **arrangement of metulae and phialides:** the entire vesicle or its upper portion may be covered with either a single layer of phialides (uniseriate) or two layers of metulae and phialides (biseriate);

- **colour and liberation pattern of conidia:** conidia also vary in shape and colour, and the presence of conidia chains may be in compact columns (columnar) or poorly defined columns (radiate).

Specifically, the most common species involved in human disease is A. fumigatus, followed by A. flavus, A. terreus and A. nidulans and A. niger. The current understanding regarding the diagnostic features of these five medically important Aspergillus species is summarised in Table 4.1. A.fumigatus were morphologically identified based on dome-shaped vesicles, uniseriate conidial heads, blue-green

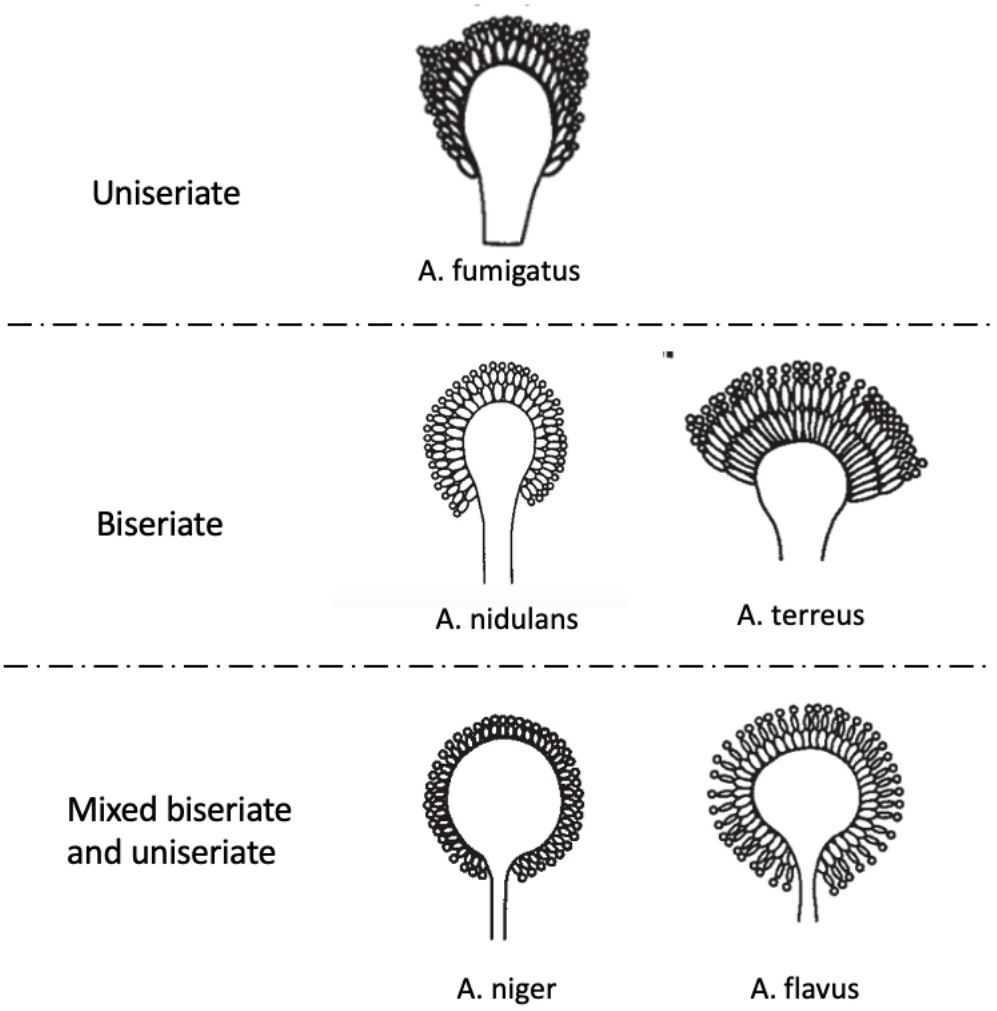


Figure 4.9: A schematic diagram of the conidial heads of different Aspergillus species.

colonies and parallel phialides in a columnar arrangement. The distinguishing microscopic characteristics of A. flavus were globose to sub-globose vesicles, mixed uniseriate and biseriate seriations, yellow to dark green conidial colour and loosely radiate arrangement. The correct identification of A. terreus was based on the dome-shaped vesicles, biseriate conidial heads, tan to brown colonies and columnar conidial arrangement. Important microscopic features of A. nidulans included semi-spherical vesicles, biseriate conidial heads, dark green colonies compactly columnar heads. A. niger was morphologically characterized by globose vesicles, biseriate conidial heads, brown to black colonies and radiate arrangement of metulae and phialides.

Species	Features of Conidial heads			
	Vesicle size (μm)	Vesicle shape	colony colour	Arrangements
A.flavus	24-45	globose	yellow to dark green	upper 3/4, loosely radiate, mixed uniseriate and biseriate
A.fumigatus	20-30	dome shaped	blue-green	upper 2/3, columnar and roughly parallel, uniseriate
A.nidulans	8-12	hemi-spherical	dark green	upper 1/2, compactly columnar and biseriate
A.terreus	10-16	dome shaped	tan to brown	upper 1/2 to 3/4, columnar, biseriate
A.niger	30-75	globose	brown to black	entirely, radiate and mostly biseriate

Table 4.1: Morphological characteristics used for Aspergillus identification.

4.3.3 Data Acquisition

The clinical data used in this chapter were acquired as part of a prospective study of anti-fungal drugs at Peking Union Hospital, China. As illustrated in Figure 4.10, the laboratory procedure for confirming fungal species includes (i) collection of clinical samples, (ii) isolation and culture for recovery of the fungal organisms, (iii) microscopic examination of properly stained samples, and (iv) informative area imaging. The sample collected depends upon the suspected location of the infection. It may include the collection of skin, nails, blood, sputum, urine, cerebrospinal fluid and/or the collection of a tissue biopsy. The medium used for the isolation should be carefully selected by clinical mycologists, the choice of which may depend on availability, cost, suspected fungal species and personal preference. In this study, the most commonly used media Sabouraud dextrose agar (SDA) was selected to support fungal growth while inhibiting bacteria. Moreover, Lactophenol cotton blue (LPCB) is commonly used to increase the visibility of the fungus in the microscopic examination of fungal cultures. LPCB is composed of phenol, lactic acid, glycerol, and aniline blue dye, with aniline blue staining the outer cell wall of the fungus and lactic acid acting as an additional clearing agent. Last but not least, the stained glass

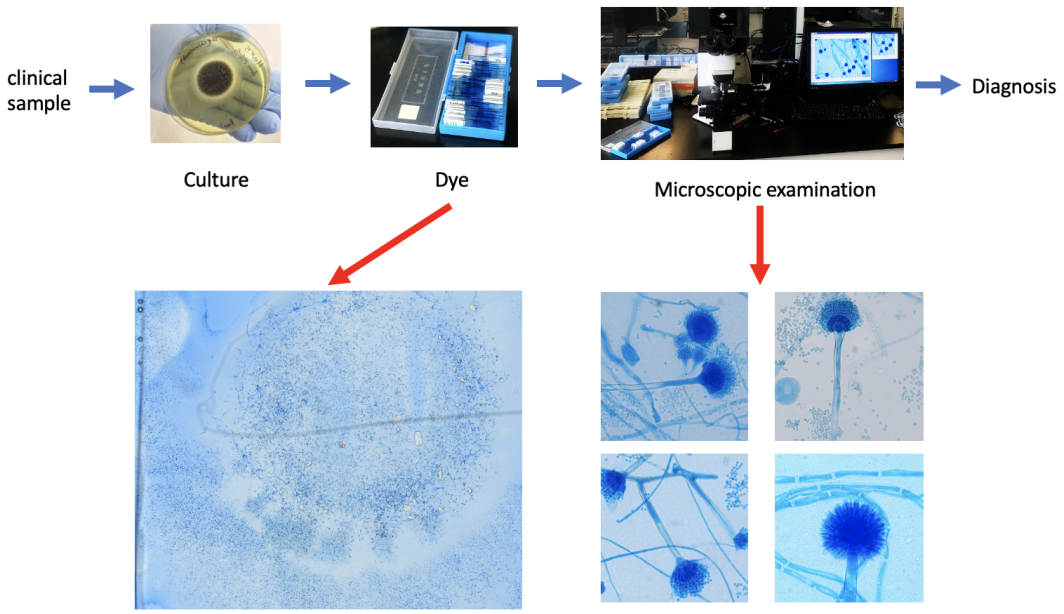


Figure 4.10: An overview of the process of data acquisition.

slides were examined directly under the microscope for fungal species identification, while the most informative regions were extracted for dataset construction.

4.3.4 Dataset Description

With the benefits of imaging and deep learning technologies, we aim to digitise the clinical samples, build a clinically relevant fungi dataset and develop a computer-aided diagnosis system to promote research on fungal infections. To the best of our knowledge, our ongoing project is the first attempt for bringing deep learning-based approaches to fungal species identification. I collaborated with Peking Union Medical College Hospital and collected 2151 microscopy images from 59 patients. The review boards at Peking Union Medical College Hospital waived ethical approval since all images were anonymised before being passed to my research group. The clinical samples are from various suspected locations of the infection in the human body. All images in this dataset have been reviewed by professional mycologists to ensure that fungal species can be objectively identified from the content of each image. In this dataset, I particularly focus on the five most common species involved in human

Species	No. of patients	No. of images
Aspergillus fumigatus	27	432
Aspergillus flavus	7	427
Aspergillus niger	8	422
Aspergillus terreus	9	424
Aspergillus nidulans	8	446

Table 4.2: Statistic information of the clinical fungi dataset.

disease: Aspergillus fumigatus, Aspergillus flavus, Aspergillus niger, Aspergillus terreus and Aspergillus nidulans. Some sample image are presented in Figure 4.11. The distribution of the clinical samples and fungal species is depicted in Table 4.2. For our experiments, the samples in each class were randomly split in a radio of $1:3$, and construct the test and training sets. In order to assess the diagnosis of fungal species in a clinically correct manner, it is important to avoid the simultaneous presence of images of the same patient in training and test sets. Thus, the split between the training set and the test set was performed at the patient level rather than on the images, meaning that there is no image from the same patient being present in both the training and testing set.

4.4 Experiments and Results

4.4.1 Implementation details

The framework was implemented using open source software library MXNet. In all experiments, input images were first resized to 224×224, and a $Resnet-18$ was used to extract global image-level information from down-sampled images. After the instance localisation, extracted patches were rescaled to 336×336, and fed to a $Resnet-50$ for final image-level prediction. Note that other CNN architectures can also be used here. To improve the training efficiency, I initialised the network with weights pre-trained on the ImageNet dataset. The whole network was trained with two NVIDIA GeForce 1080 Ti GPUs and the mini batch size was set to 16. I used the stochastic gradient descent (SGD) optimiser with an initial learning rate of 0.05 that was multiplied by 0.1 after every 50 epochs. The initial weight score λ in the loss function is 1, and reduced by 0.1 after every 20 training epochs.

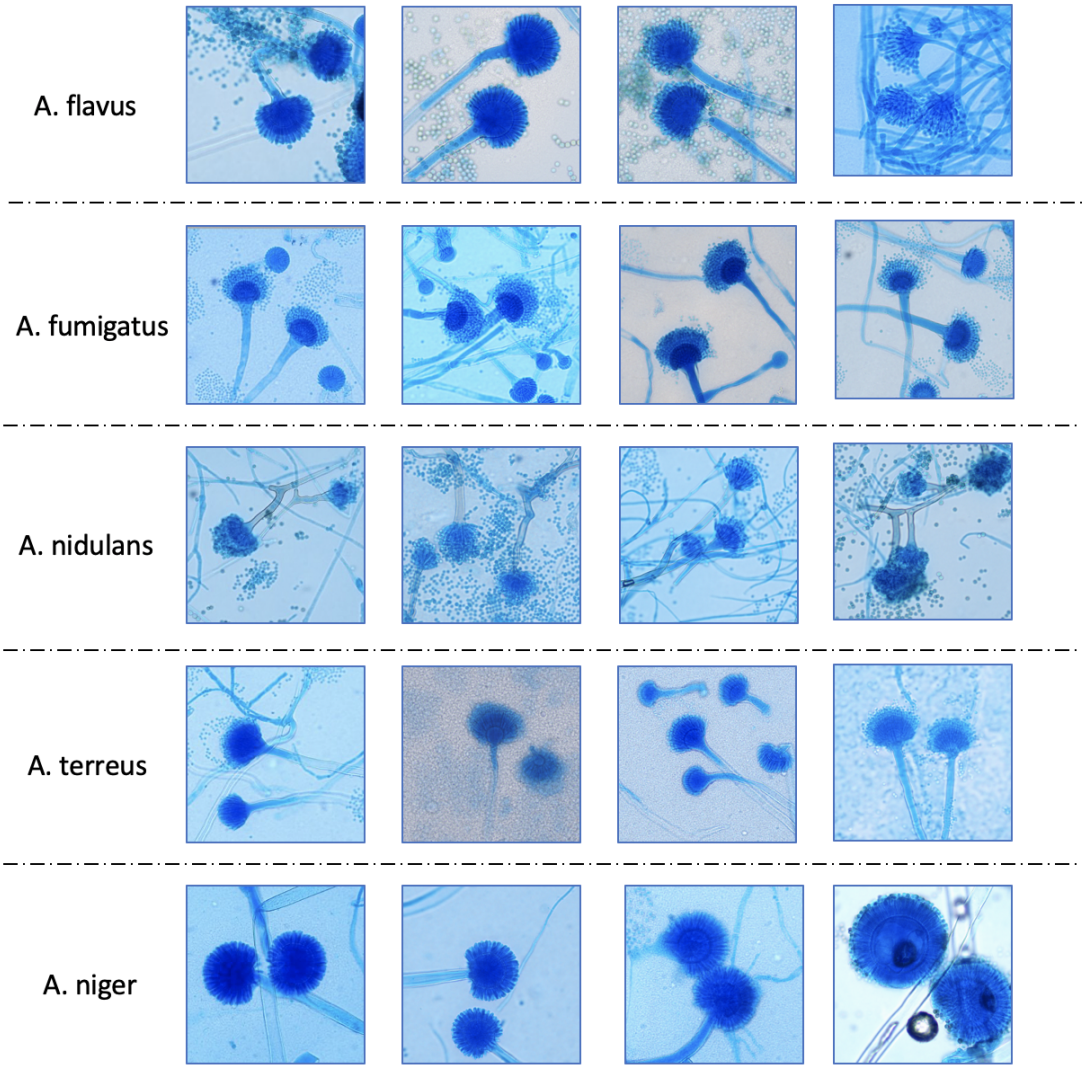


Figure 4.11: Examples for the five classes in the Fungi dataset.

4.4.2 Quantitative Comparison with Competing Methods

This section evaluated the effectiveness of the proposed method by comparing it with several state-of-the-art classification methods. Table 4.3 presents the classification accuracies. All of the compared methods were trained with the same backbone and computing environment. This section first shows the baseline performance of backbone network Resnet-50 [71], and describes experimental results on two attention-based techniques that can be easily plugged into the Resnet-50. The Residual Attention Network (RAN) [113] was built by stacking multiple

Method	top-1 % accuracy
Resnet-50 [71]	0.907
Residual Attention Network[113]	0.867
Attend and Rectify [112]	0.871
Trilinear Attention Module[119]	0.883
NTS Module [153]	0.914
Our Method	**0.943**

Table 4.3: Experimental results in Fungal dataset.

attention modules which employed the bottom-up top-down structure to generate the attention-aware features highlighting the informative regions on convolutional feature maps. Inspired by the original Resnet, they proposed a residual learning mechanism for their stacking attention modules to avoid performance degradation as the number of layers deepens. Similarly, Rodriguez et al. [112] proposed another universal architecture-independent attention module that can be easily incorporated into any CNN-based architecture to emphasise discriminative features for classification. The proposed attention module consisted of three main sub-modules: (i) the attention heads, which defined the most relevant regions, (ii) the output heads, which generated class probability scores, and (iii) the confidence gates, which outputted a confidence score for each attention head. Although such a modular design is fast, simple and independent of architecture, it can be observed that the performance of these two stacked networks is not close to the state-of-the-art. One possible reason is that these methods failed to eliminate the effect of noisy background and attended to some irrelevant features when the number of training samples was limited. Moreover, the classification performance may be highly sensitive to the choice of hyperparameters, i.e., the number of attention heads in [112]. Additionally, these methods only extracted global features from down-scaled low-resolution images, and the direct enhancement in the high-level feature space made the networks lack high-resolution information, which may result in sub-optimal results in biomedical applications.

This section also compared our method with other two-stage classification networks which required a sampling operation to extract the most discriminative

regions. Zheng et al. [119] designed a trilinear attention network that used a channel-wise attention mechanism to localise the discriminative parts and employed an attention-based soft sampling scheme to resample patches from the original image. As the discussion before, soft sampling strategies can hardly guarantee sufficient shape information of multiple biomedical objects, leading to dropped performance. In addition, the attention-based sampler generated a local informative patch by randomly selecting a channel for each iteration in the training process. I suspect that in the absence of a sufficiently large training set, it is difficult for the network to ensure that each channel is aware of meaningful features, which is why I added all channels together along the depth direction to generate the attention map. Yang et al. [153] also proposed a two-stage pipeline which was trained with reinforcement training. Specifically, their Navigator-Teacher-Scrutiniser(NTS) network consists of a Navigator agent, a Teacher network, and a Scrutiniser network. The Navigator generated a bunch of region proposals for each input image, and produced a list denoting the informativeness of all anchors. The regions with the highest amount of information were cropped and fed into the Teacher network to obtain confidence scores. In the scrutiniser network, they concatenated all local features for final image-level classification. I slightly modified the original implementation to use a smaller anchor box to capture the vesicles in the fungi images. The resulting cooperative learning scheme achieved the promising classification performance, but still heavily relied on pre-defined region proposal scubas the scales and ratios of the anchors. In addition, reinforcement learning made the network slow to train. The proposed method follows Yang et al.'s idea of concatenating multiple local information patches. However, rather than using reinforcement learning, our method incorporated the attention mechanism to directly attend to multiple informative regions. From the comparison shown in Table 4.3, we can observe that our multi-instance attention-based network achieved the best performance over other classification methods, with a 2.9% improvement over NTS network.

4.4.3 Comparative Evaluation of Attention Mechanisms

In order to analyse the influence of the proposed attention module, I also conducted experiments with various attention mechanisms. I used the same sampling strategy, feature fusion scheme and loss function in all experiments when comparing the attention mechanisms. Specifically, I only modified the attention-based instance localisation module in our framework, which enabled us to rigorously evaluate the localisation capability of attention modules. Table 4.4 shows the results of integrating top attention modules in our classification framework.

I compared the proposed model to recent state-of-the-art attention modules used in computer vision and biomedical image analysis. Wang et al. [117] presented a widely-used non-local operation that computed the attention weight of the query position based on the pixel-level pairwise relationships between this query position and all possible positions in the feature maps. This allowed the network to capture the long-range contextual information between distant pixels. Instead of capturing positional correspondence, Zheng et al. [119] extended this idea to explicitly model the inter-channel correspondences. They first obtained bilinear features that represent the spatial correspondence between channels and further integrated the spatial relationships into the feature map by performing dot production. Dual attention module, proposed by Fu et al. [110] combined the above ideas to capture features dependencies in both spatial and channel dimensions. They built two parallel attention modules: one is a position attention module and the other is for channel attention, where the outputs were concatenated to further strengthen the feature representations.

Apart from the attention modules based on pairwise computation, I also compared the proposed attention mechanism with two methods where fully-connected layers were used to learn the attention weights. The Squeeze-and-Excitation block [114] adaptively recalibrated channel-wise feature responses by explicitly modelling interdependencies between channels. The first step applied the global average pooling to squeeze the feature maps into one-dimensional channel descriptors. In the second step, a fully connected neural network with two hidden layers was used

Attention Mechanism	top-1 % accuracy
Non-local spatial attention [117]	0.859
Trilinear channel attention [119]	0.883
Dual attention [110]	0.901
Squeeze-Excitation attention [114]	0.939
Global-context attention [120]	0.937
Our Method	**0.943**

Table 4.4: Comparison of global attention mechanisms.

to automatically learn channel-wise nonlinear interactions. Following this line of research, Cao et al. [120] designed a three-step attention module. (a) a context modelling module, which uses a global average pool to aggregate features together; (b) a bottleneck transformation module, which uses a 1x1 convolution, a ReLU, and a 1x1 convolution to capture channel aspect interdependencies; and (c) a fusion function, which merges global context features into the original features. In contrast to the soft-attention mechanism described above, this chapter presented a novel attention module that directly filtered out the uninformative feature maps and preserved the shape information for fungal vesicles. As reported in Table 4.4, the proposed attention module outperforms all other attention-based algorithms, demonstrating the effectiveness and localisation ability of the proposed method.

4.4.4 Visualisation of the Attention Map

In order to demonstrate the effectiveness of the proposed attention mechanism, the attention maps of some sample images in the test set are shown in Figure 4.12. As a result of the multi-instance localisation module, the attention maps depict which instances the network pays attention to, and where they are localised. Note that the first column is sample images of each class in our dataset, and the second column is the visualisation result of the last convolutional layer in a standard Resnet-50 [71], while the last column shows the corresponding attention maps generated in the proposed network. We can see that the standard CNN model is insufficiently powerful for localising semantically meaningful regions, whereas our attention-based localisation module tends to accurately localise the conidial

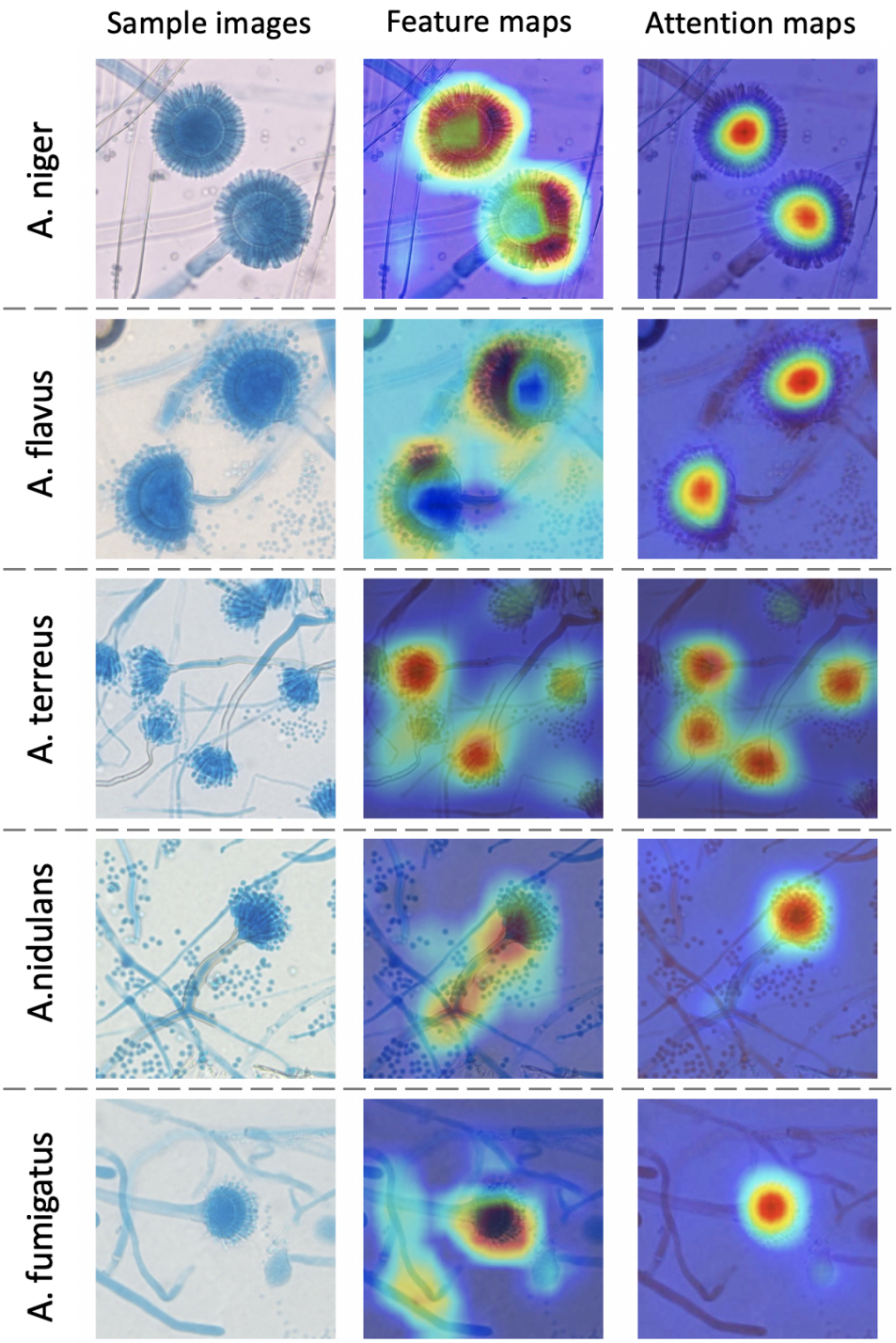


Figure 4.12: Visualisation of global attention maps. (a) Each row shows a sample image from each class represented in the fungal dataset. Our attention maps suppress features associated with background clutters and focus on the whole objects.

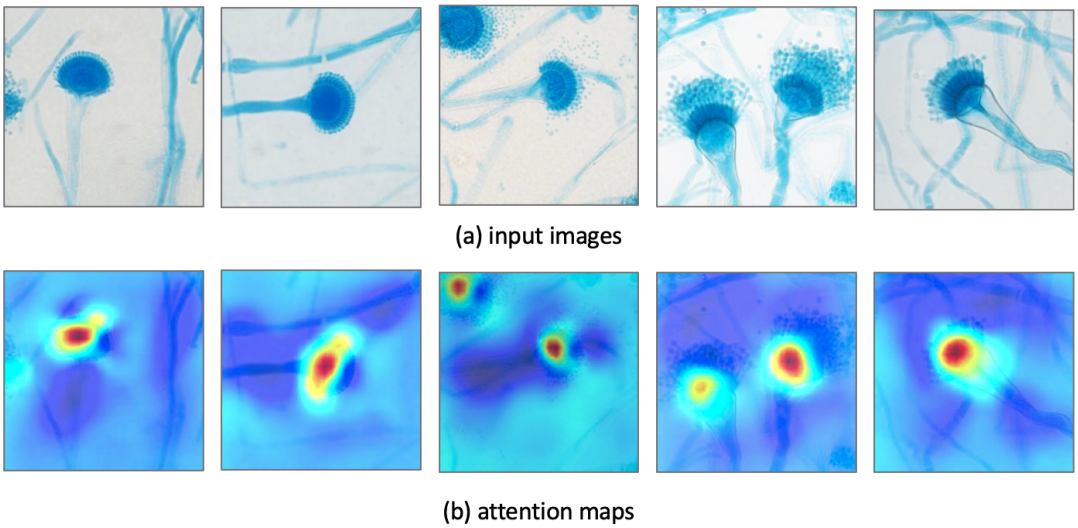


Figure 4.13: Failure cases of global attention maps within the test set.

regions from the entire image for the future image-level prediction, even though these regions are dispersed and some of them are small.

The failure cases of the attention module are reported in Figure 4.13. It can be seen that the proposed framework fails to effectively suppress the background clusters when there is a high degree of variation with respect to orientation, illumination, background or object appearance. One possible reason for this is that we selected the top 100 information channels in the channel filter module for all images in the dataset to construct the final attentional maps. For these difficult cases, some confounding channels may also be included, and thus the network fails to suppress irrelevant information and focuses attention on the vesicle regions. In the next chapter, a potential improvement will be carried out to cope with these failed instances using a gated attention mechanism. In addition, a weighted feature fusion procedure will be introduced to effectively reduce the impact of these non-informative instances. This will further direct the network to pay attention to the most discriminative instances and ignore irrelevant information.

4.5 Conclusion

Unlike general image recognition, many real-life applications may involve the classification of subclasses belonging to the same basic class with limited data, such as species identification in biomedical image analysis. Recently a large number of fine-grained classification methods have been proposed for natural images. They successfully apply machine learning or deep learning techniques to automatically analyse the content of images, such as object or discriminant part localisation. However, this task remains challenging for existing methods, as they need to be capable of locating and distinguishing subtle visual distinctions in local discriminative regions within subordinate categories. In particular, biomedical images may contain multiple discriminative instances, while it is important to localise these instances simultaneously so that both global structural information and local instance-level features can be captured. This has prompted the need to advance state-of-the-art fine-grained classification methods for biomedical images.

This chapter presents a novel attention-based multi-instance classification framework that enables the joint localisation of discriminative instances and suppresses confounding effects in irrelevant regions even if the size of the dataset is limited. High-level structural information and local details of key instances are aggregated for the final image-level classification, but no additional annotations or region proposals are required. The experimental results demonstrate the promise of the proposed method: the quantitative results establish a new state-of-the-art technique for fungi species classification, while the qualitative results demonstrate the localisation capabilities of the proposed attention-based module. The time and complexity analysis will be provided in Chapter 5. As part of this work, I also constructed a new clinical fungi dataset comprising of five species that most frequently affect humans. To the best of our knowledge, this project is the first attempt for bringing deep learning-based approaches to fungi species recognition.

Although great progress has been reported, there are still several limitations to this design. First, the computational cost of the proposed attention module is relatively high. Then the high intra-class variations and low inter-class variations

make it challenging to extract consistent fine-grained features for instances within the same label. More importantly, this method tends to deal with the detected instances independently, ignoring their intrinsic dependencies, and thus lack the ability to regulate the contribution of each instance to the final image-level classification. In the next chapter, three technical extensions will be presented to address these limitations. In addition to fungal species identification, the next chapter will also further explore how the proposed framework addresses different biomedical image classification tasks.

5

An Interpretable Gated Attention Mechanism

Contents

5.1 Introduction 116
5.2 Method 119
5.2.1 First Extension: Gated Attention Mechanism 119
5.2.2 Second Extension: Weighted Feature Fusion Scheme . . 121
5.2.3 Third Extension: Instance Representation Loss 122
5.3 Datasets 124
5.3.1 Clinical Fungi Dataset 124
5.3.2 BreAst Cancer Histology images (BACH) 125
5.3.3 ISIC-2017 Dataset 125
5.3.4 Coeliac Disease Dataset 127
5.4 Experiments and Results 131
5.4.1 Implementation details 131
5.4.2 Evaluation and Analysis on Clinical Fungi Dataset . . . 131
5.4.3 Evaluation and Analysis on BACH dataset 139
5.4.4 Evaluation and analysis on ISIC-2017 Dataset 139
5.4.5 Evaluation and Analysis on Coeliac Disease Dataset . . 142
5.5 Clinical Background Alignment 144
5.5.1 Breast Cancer Classification from Histopathology Images 145
5.5.2 Skin Cancer Classification from Dermatoscopic Images . 147
5.5.3 Fungal Species Classification from Mycology Images . . 148
5.5.4 Coeliac Disease Classification from Histopathological Images 150
5.6 Conclusion 151

As discussed in Chapter 4, there are still some limitations to this attention-

based multi-instance classification framework. This chapter aims to address these issues by proposing three technical extensions to the framework. This chapter also demonstrates the superiority of the proposed fine-grained classification pipeline by evaluating it on four different biomedical datasets. The quantitative and qualitative results show that it not only achieves state-of-the-art classification performance, but also consistently learns clinically relevant features for decision-making.

Originality: This chapter used a new coeliac disease dataset. The whole-slide images and the ground truth were provided by collaborator Dr. Holm Uhlig. For preprocessing, I independently implemented a convolutional autoencoder and k-means algorithm to extract the patches around the villous regions. I also independently designed and implemented three technical extensions in Python and MXNet. Finally, I independently conducted validation experiments and performed visual comparisons of the generated attention maps and clinically relevant features. The first technical extension presented in this chapter has been reported in:

Mengran Fan, Tapabrata Chakraborti, Eric I-Chao Chang, Yan Xu, Jens Rittscher. Microscopic fine-grained instance classification through deep attention. *International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI)*, 2020.

5.1 Introduction

A novel attention-based multi-instance framework for fine-grained classification of fungi species has been presented in Chapter 4. The pipeline enables the joint localisation of multiple discriminative instances and explored the local high-resolution and global structural features simultaneously. Despite the fact that the algorithm takes into account that biomedical images require simultaneous extraction of features from multiple discriminative instances and has made great progress in the task of fungal species identification, there are a number of limitations in the design of the network.

First, the proposed attention module consists of three different sub-modules, resulting in a high cost of time and computational resources. In addition, the

proposed channel filter requires a hyperparameter to determine how many channels will be selected to participate in the subsequent transformations, which may need to be fine-tuned when applied to other datasets. In this chapter, I introduce an interpretable gated attention mechanism that uses a gated mechanism to eliminate the confounding effects of irrelevant channels. In contrast to the channel filters proposed in Chapter 4, these gated layers can be viewed as a soft filter that forces the model to pay more attention to information channels in a way that has limited computational cost and no hyperparameters.

The second limitation is that the proposed approach tends to treat the detected instances independently, ignoring their intrinsic dependencies, and thus lacks the ability to moderate the contribution of each instance to the final image-level classification. In addition, how to meaningfully aggregate local information, rather than simply concatenating instance-level features, still requires further investigation. To this end, this chapter proposes a weighted feature fusion scheme that dynamically recalibrates the weights of sampled instances for final image-level predictions. The number of patches sampled from high-resolution data is fixed across the whole dataset, therefore when there are a small number of discriminative instances in the image, this strategy may face a risk that not all sampled patches are meaningful for the final fine-grained predictions. In this case, the proposed weighted fusion module can effectively highlight the most discriminative instances and suppress the contribution of less informative instances.

The high intra-class variation and low inter-class variation in biomedical images make it challenging to extract consistent fine-grained features for instances within the same label. This chapter also aims to improve the consistency and discrimination of the learned patch-level feature representations. This chapter discusses an instance representation loss to potentially minimise within-class variance at the instance level. In order to investigate the effectiveness of these three extensions, an ablation study of different components in this proposed pipeline is also presented. The classification performance is evaluated on four different biomedical classification datasets, thus demonstrating that the method is a powerful general approach to

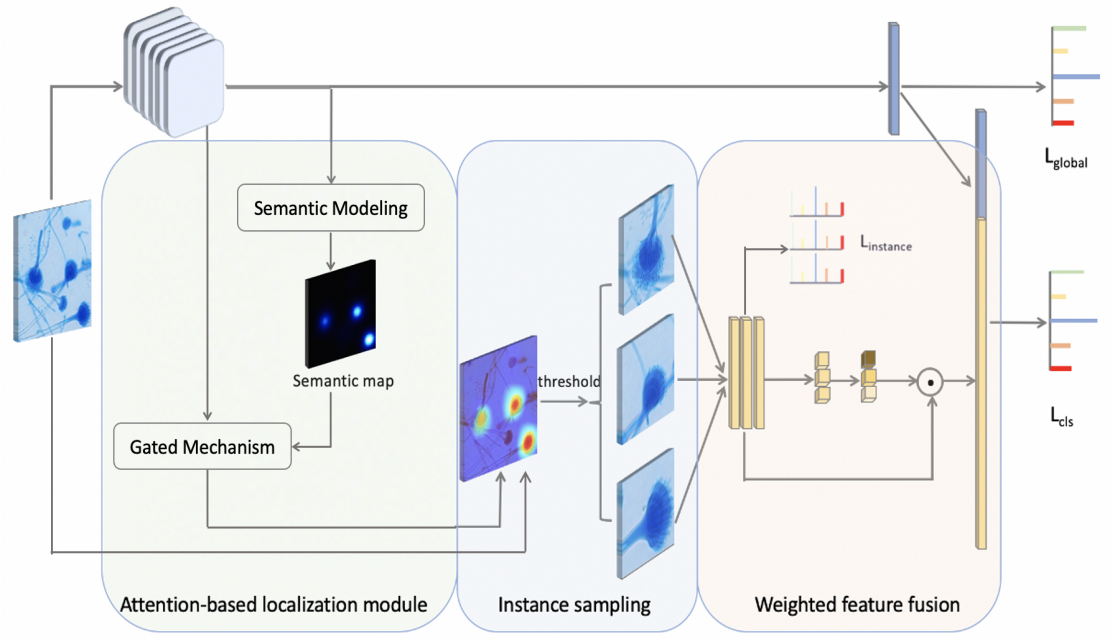


Figure 5.1: Framework for the proposed attention-based fine-grained classification pipeline. In the attention-based localisation module (Section 5.2.1), multiple instances are highlighted based on global image-level features. Details of the proposed semantic modelling and gated mechanism are depicted in Figure 5.2 and Figure 5.3, respectively. However, the number of extracted instances is fixed across the whole dataset, and when the number of discriminative instances in the image is small, this strategy may face a risk that not all sampled instances are meaningful for the final fine-grained prediction. Therefore, a weighted feature fusion scheme (Section 5.2.2) and an instance representation loss (Section 5.2.3) are introduced to enhance the instance-level feature learning.

achieve superior performance in different biomedical scenarios. This chapter also contributes to the utility and reliability of deep learning techniques in clinical practice by demonstrating in details the strong agreement between learned features and medical knowledge, demonstrating that our approach can be applied as an interpretable model to a wide range of biomedical image classification problems.

The technical extensions are described in Section 5.2. Then the datasets used to assess the performance of the proposed pipeline are introduced in Section 5.3. Finally, the details of experimental results and clinical alignments are presented in Section 5.4 and Section 5.5, respectively.

5.2 Method

5.2.1 First Extension: Gated Attention Mechanism

In natural images, noisy background or irrelevant regions are highly variable and can be therefore naturally discarded by deep neural networks. However, for biomedical datasets with limited training samples, this is more difficult to achieve. Here, this chapter introduces a gated attention mechanism that is aimed to filter out the confounding channels and localise the most discriminative instances without extra part supervision or redundant region proposals. As shown in Figure 5.1, the module first summarises a $2D$ semantic map $S \in \mathbb{R}^{H \times W}$ from the input convolutional feature maps $X \in \mathbb{R}^{H \times W \times D}$. Furthermore, the method uses the high-level semantic information to filter out the confounding channels based on the pairwise correspondences between each input channel and the generated semantic map, thus suppressing the irrelevant background and localising the most discriminative instances.

Semantic Modelling. Since spatial attention weights computed for each query position are almost the same for different tasks [120], a global spatial attention map is extracted from the input features, which is shared by all query positions within an image. For multi-level semantics understanding, two dilated convolution blocks [166] were first applied to the output of the feature extractor X, as illustrated in Figure 5.2. The set of multi-scale features are compressed by computing the sum of all channels $M = \sum_{k=1}^{D} W^T X$ where W^T is the weights of the dilated blocks. This channel compression rests on the assumption that if the region is activated on most channels, the region tends to be more discriminative and to have a higher likelihood of being part of the object of interest. The final semantic map S is generated by applying a spatial softmax layer that performs the softmax operation over all feature points in the aggregated map M, resulting in a probability distribution that roughly indicates the regions of the most discriminative instances:

$$S_{i,j} = \frac{\exp\left(M_{i,j}\right)}{\sum_{l=1}^{H} \sum_{k=1}^{W} \exp\left(M(l,k)\right)} \tag{5.1}$$

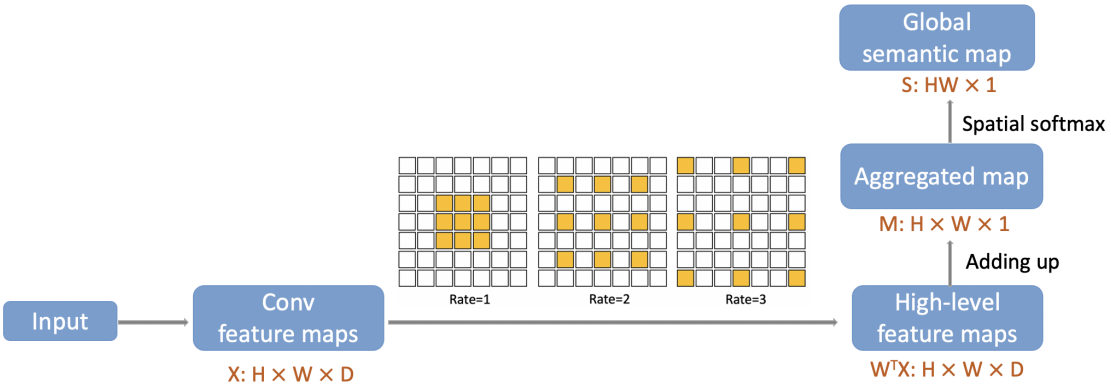


Figure 5.2: Semantic modelling. In semantic modelling, two dilated convolution blocks [166] were first applied to the convolutional feature vector $X \in \mathbb{R}^{H \times W \times D}$, where H, W, D stand for the height, width and channels respectively. Then the aggregated map $M \in \mathbb{R}^{H \times W \times 1}$ is obtained by computing the sum of all channels in high-level feature maps: $M = \sum_{k=1}^{D} W^T X$ where W^T is the weights of the dilated blocks. Finally, the high-level semantic map is obtained by applying a spatial softmax layer.

Gated Mechanism. To measure the discriminability of each channel, the proposed method captures the spatial correspondence scores via conducting matrix multiplication over X^T and S, where X^T is the original convolutional feature maps with the shape of $D \times HW$ (Figure 5.3). For example, X_k is the k^{th} channel of the input feature maps, containing its specific semantic responses. So $X_k^T S$ is the importance coefficient that indicates the semantic representation power of this channel. Our method is different from traditional channel-based self-attention mechanisms [110, 119] that usually directly capture the pairwise inter-channel dependencies by calculating $X^T X$. In order to enhance the specific semantics, this method summarises a global high-level semantic map and uses it as a template to quantify the representation capability of each channel by $X^T S$. Therefore, such a mechanism were applied to obtain the $1D$ coefficient vector $X^T S \in \mathbb{R}^{D \times 1}$ rather than $X^T X \in \mathbb{R}^{D \times D}$.

To filter out the confounding channels, the hyperbolic tangent ($tanh$) and ReLU activation functions are used to normalise the discriminability coefficient among all channels. As a result, a set of gated weights is obtained, selects channels that look at the most discriminative regions. In particular, the gated weight is approximately 1 for the most informative channels, and approximately 0 for the

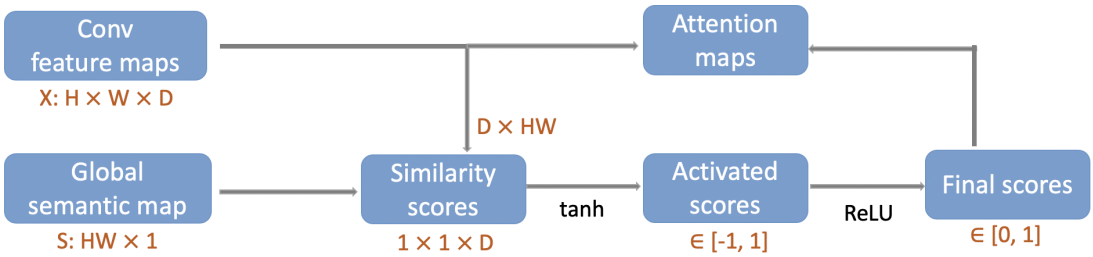


Figure 5.3: Gated Mechanism. To measure the discriminability of each channel, the proposed gated mechanism captures the spatial similarity scores between the convolutional feature vector $X \in \mathbb{R}^{H\times W\times D}$ and the high-level semantic map $S \in \mathbb{R}^{HW\times 1}$, where H, W, D stand for the height, width and channels respectively. The hyperbolic tangent ($tanh$) and ReLU functions are used as gated activation layers to filter out the confounding channels. Finally, the attention map is modelled as a gated average of the outputs of the original convolutional feature maps.

channels highlighting the irrelevant background. The gated activation layer can also be regarded as a filter which enforces the model to ignore the confounding channels and pay attention to more informative channels. Consequently, to let the attention module focus on multiple instances, the final attention map is modelled as a gated average of the outputs of the original channels.

$$\Omega_{att} = \frac{1}{D}\sum_{k=1}^{D}(X_k \odot ReLU(\tanh\left(X_k^T \otimes S\right))) \tag{5.2}$$

5.2.2 Second Extension: Weighted Feature Fusion Scheme

The framework shown in the previous chapter tends to treat the detected instances independently, ignoring their intrinsic dependencies and making it difficult to estimate the contribution of each instance. Therefore, it is difficult to make accurate image-level predictions without taking into account interrelated information across instances. To this end, this section introduces a weighted feature fusion scheme to explore the special contributions of instances for improving classification performance. Given the instance-level feature embeddings $H = \{h_1, h_2, \ldots, h_N\}$, the proposed fusion module assigns importance weights to each instance to highlight the discriminative instances and to suppress the irrelevant instances within the image. In particular, the importance weight α_i for each instance is determined by

a softmax function, where the weights are normalised into a discrete probability distribution indicating the importance coefficient over all instances (the summation over all coefficients is 1):

$$\alpha_i = \frac{\exp\left(\delta(h_i)\right)}{\sum_{j=1}^{N} \exp\left(\delta(j_i)\right)} \tag{5.3}$$

where $\delta(.)$ denotes the global average pooling that squeezes the instance-wise embeddings into one dimension. Then the new feature representation of the i^{th} instance can be formulated as: $h^{'} = \alpha_i h_i$. Finally, the method concatenates all obtained feature vectors $h_i^{'}$ and consolidates these weighted instance-level features with global image-level features for the final predictions.

Our lightweight fusion module does not introduce any additional trainable parameters compared to the soft-attention mechanisms proposed in other multi-instance related work. In our gated attention module, meaningful instances are highlighted based on the global image-level features. However, the number of extracted instances N is fixed among all images, which limits the flexibility and discriminative power of the instance-level feature extraction. Hence, this method aims to use a lightweight soft-attention mechanism to further refine the instance-wise representation based on local features extracted from individual instances. The entire design can be considered as a two-stage multi-instance system: the gated attention module is used to identify discriminative instances from global image-level features, and the weighted fusion module is used to highlight strongly activated instances based on local instance-level dependence, as illustrated in Figure 5.1. In summary, the weighted feature fusion scheme is lightweight and sufficiently powerful to recalibrate the highlighted instances in the gate attention module.

5.2.3 Third Extension: Instance Representation Loss

According to Figure 5.1, the weighted feature fusion module jointly provides an instance vector and a corresponding importance coefficient for each instance. Intuitively, instances extracted from the same image should be similar, as they have the same image label. Due to large intra-class variation, however, it is quite

challenging to extract consistent fine-grained features at the instance level despite the fact that patches of instances from the same image have the same labels. To address this problem, this section proposes an instance representation loss $L_{instance}$ that enhances the intrinsic consistency between the most discriminative instance and all other sampled instances. Inspired by the knowledge distillation strategy [167], a fully-connected layer and a softmax layer were first applied to each instance to generate a set of probability vectors p over the categories. Based on these probability vectors, the proposed loss function employs soft-target cross entropy to force the model to learn consistent discriminative features for instances from the same image, potentially minimising the intra-class variance:

$$L_{instance} = -\sum_{i=1}^{N}\sum_{k=1}^{M} p_i^{(k)} \log p_{key}^{(k)} \tag{5.4}$$

where M is the number of classes, and N indicates the number of sampled instances. p_{key} is the predicted probability of the instance that is assigned the highest importance coefficient. The purpose of this loss is to enable the model to learn consistent features across the most and least recognisable instances, thereby enhancing the discriminatory power of instance-based feature learning.

Therefore, I designed a network with three types of loss functions, namely, global attention loss, classification loss, and instance representation loss. Specifically, the entire network is optimised through a multitasking loss function, which is defined as follows.

$$\mathcal{L} = \lambda L_{global}(Y_g, Y^\star) + (1-\lambda)(L_{fusion}(Y_f, Y^\star) + L_{instance}) \tag{5.5}$$

where L_{global} and L_{fusion} are standard cross entropy losses with respect to the outputs of the global image-level network Y_g and the proposed multi-instance fusion network Y_f, respectively. $Y^\star$ represents the ground truth label and λ is a weight parameter that is initialised as 1 and gradually decreased during the training process. As discussed in Section 4.2.3, there are two levels of deeply learned features in the proposed multi-task learning scheme: Y_g optimises the global context-aware

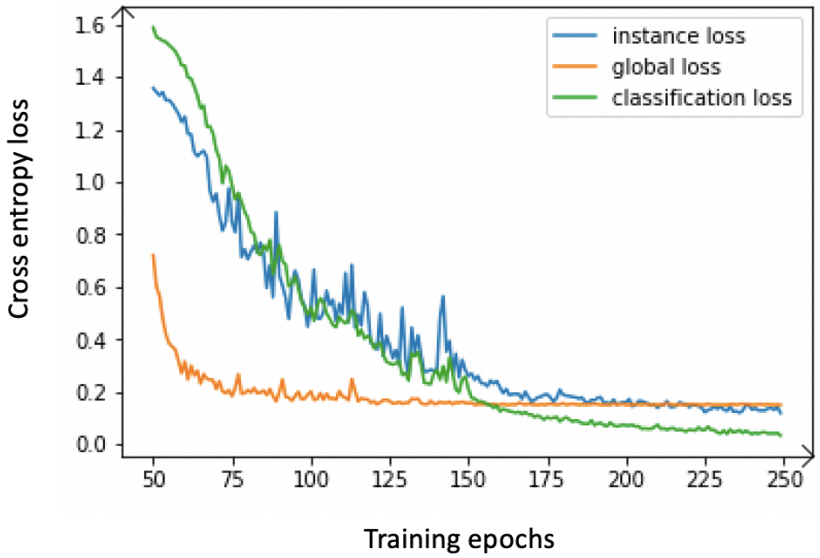


Figure 5.4: Line plots of the different loss term in Eq. 5.5 in the joint learning process.

features, while L_{fusion} and $L_{instance}$ enhance discriminative and consistent instance-level feature learning. The loss of the global image-level network converges faster, which allows the network to focus on extracting global high-level semantic features at the beginning and gradually increase the contribution of instance-level features during the training process. Therefore, the discriminative power of the fine-grained feature learning can be greatly improved. To further justify the form of Eq. 5.5, line plots of the different loss term in the joint learning process are depicted in Figure 5.4, which indicates L_{fusion} and $L_{instance}$ converge at similar rates.

5.3 Datasets

The proposed classification framework were evaluated on four different biomedical image datasets: Clinical Fungi Dataset, BreAst Cancer Histology images (BACH), ISIC 2017 skin dataset and a newly constructed coeliac disease dataset.

5.3.1 Clinical Fungi Dataset

As described in Section 4.3, in collaboration with Peking Union Medical College Hospital, I collected 2151 microscopic images from 59 patients. These samples

were taken from various suspected locations of human infection and annotated by a professional mycologist. In this dataset, I focused specifically on the five most common species involved in human disease: Aspergillus fumigatus, Aspergillus flavus, Aspergillus niger, Aspergillus terreus, and Aspergillus niger. Please refer to Section 4.3 for more details on the collection methods and experimental setup.

5.3.2 BreAst Cancer Histology images (BACH)

As part of the ICIAR 2018 conference (15^{th} International Conference on Image Analysis and Recognition), the grand challenge on BreAst Cancer Histology images (BACH) was organised to promote further research on breast cancer images and tools to reduce the diagnostic burden on pathologists. Challenge organisers provided a new annotated dataset of Hematoxylin and Eosin-stained microscopic images for classification and localisation of clinically relevant histopathological categories. Specifically, this dataset is composed of 400 high-resolution (2018×1356 pixels) microscopic images that share the same acquisition conditions and are uniformly distributed across the four classes. Images were labelled by two pathologists, and each image was assigned to one of four classes based on the predominant tissue type; 1) normal, 2) benign, 3) carcinoma in situ, and 4) invasive carcinoma (Figure 5.5). I randomly perform a 75% - 25% split for training and test sets.

5.3.3 ISIC-2017 Dataset

In 2017, the International Skin Imaging Collaboration (ISIC) hosted a grand challenge: Skin Lesion Analysis Towards Melanoma Detection Challenge. This challenge provided a large-scale publicly available dataset of skin dermoscopy images that can serve as a benchmark for melanoma diagnosis. The skin lesion classification dataset consists of 2000 training, 150 validation and 600 testing images, paired with the corresponding ground truth annotations, i.e., melanoma, nevus, and seborrheic keratosis (melanoma vs. others). Sample images from ISIC-2017 dataset are depicted in Figure 5.6. As an additional preprocessing step, all dermoscopy images are normalised by the grey-world colour constancy algorithm [168].

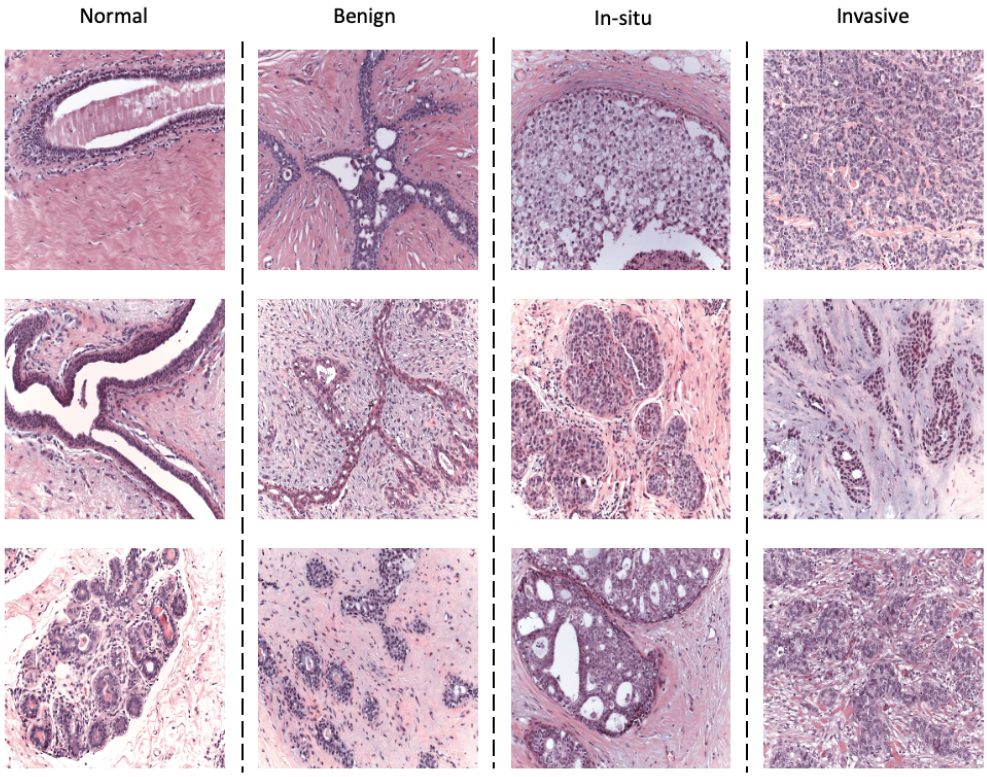


Figure 5.5: Sample images from BACH dataset.

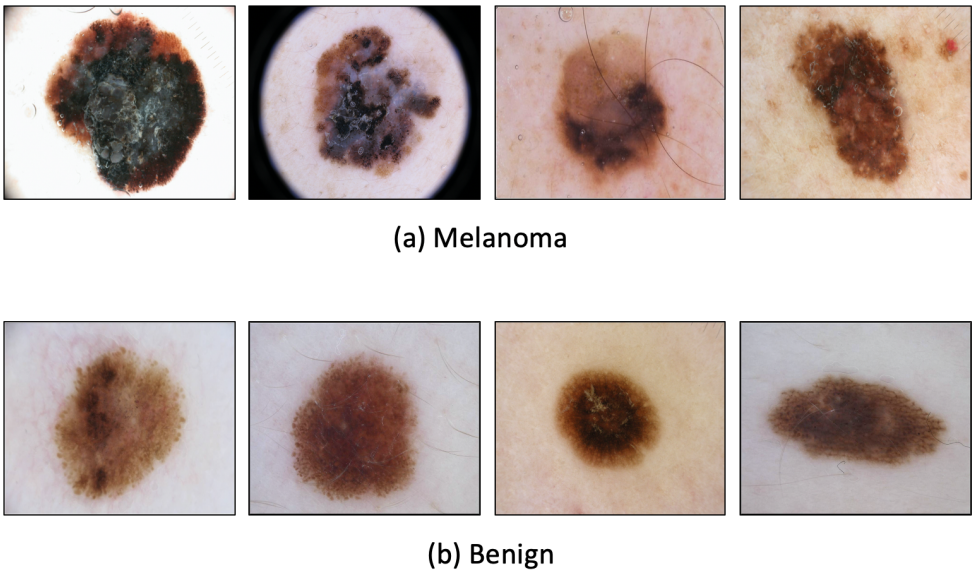


Figure 5.6: Sample images from ISIC-2017 dataset.

5.3.4 Coeliac Disease Dataset

Clinical Background. Coeliac disease (CD) is a chronic inflammatory disorder of the small intestine caused by an immune response that attacks the small intestine after ingestion of dietary gluten products (e.g. wheat, barley, rye) in susceptible individuals. Compared to patients with other autoimmune diseases, patients diagnosed with CD can be completely prevented from serious and even life-threatening complications through early serological diagnosis and a life-long gluten-free diet, which requires extensive patient education and monitoring [169, 170]. Nowadays, CD is one of the most common genetic disorders in humans, and recent studies have found that the United States and Europe are considered to be geographic areas of high prevalence, with an incidence of 1-2% [171]. The current diagnosis of CD consists of an initial specific antibody serologic test and a confirmatory small intestinal biopsy showing histologic evidence of CD. A small intestinal biopsy obtained during endoscopy is considered the "gold standard" in the diagnosis of CD. Despite advances in diagnosis, the majority of patients remain undiagnosed due to unclear clinical symptoms, which may lead to serious consequences for children and adults [171, 172]. As a result, there has been significant clinical interest in developing more efficient and accurate automatic methods for histological analysis of small intestinal biopsy images.

Histological Features. Patients may come in for a small bowel biopsy due to clinical suspicion or a positive serologic test. Typically, multiple samples are collected from the patient through an endoscopic procedure and then visually examined by a pathologist under a microscope (Figure 5.7). In particular, tissue samples collected from patients with coeliac disease have three histological changes: (1) increased intraepithelial lymphocytes (IELs), (2) crypt hyperplasia, and (3) atrophy of the villi [169, 171, 172]. Lymphocytes are a sub-type of white blood cells that are part of the immune system's response to inflammation; therefore, an increased number of lymphocytes is not specific to CD and may be seen in a variety of autoimmune disorders. A large number of subjects with CD have more

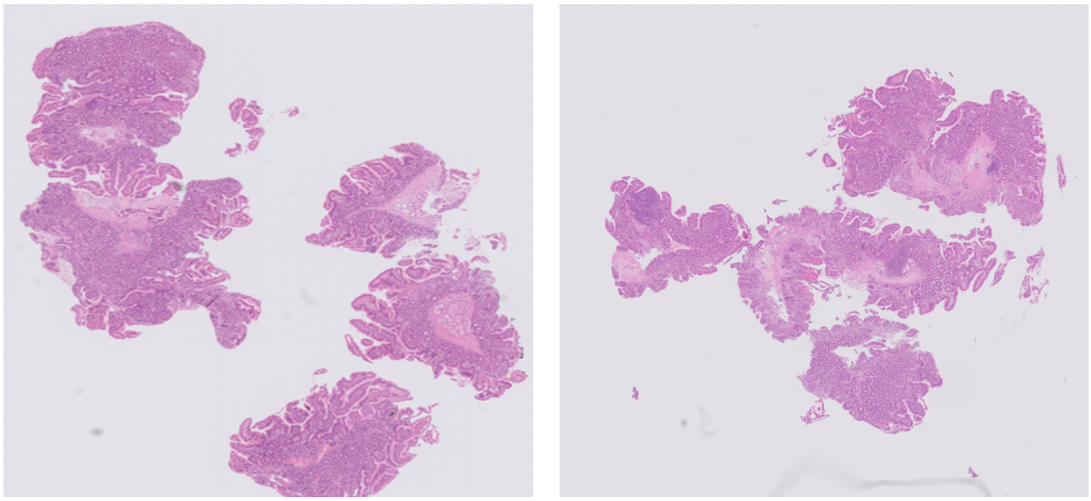


Figure 5.7: Sample whole-slide image from coleiac disease dataset.

than 30 lymphocytes per 100 epithelial cells. The villi are finger-like extrusions in the mucosa that contribute to the increase in surface area of the mucosa, and the crypts are glands located between the villi. The normal ratio of the height of a typical villus to the depth of a representative crypt should be between 3:1 and 5:1. In diseased samples, the villi are compressed and the crypts are deformed, which significantly reduces the surface area available for absorption.

Based on histological characteristics, the severity of CD is best assessed using a grading system designed by Marsh [173] and modified by Oberhuber [174] (Figure 5.8). In the Marsh-Oberhuber grading system, type 1 lesions are characterised by normal villous structure and an increased number of intraepithelial lymphocytes, which can be seen in patients with latent and occult CD, but are relatively nonspecific. Type 2 are hyperplastic lesions, featuring an increase in intraepithelial lymphocytes and crypt hyperplasia (increased depth), with the ratio of villi height to crypt depth dropping below 3:1. Type 3 are destructive lesions that can be grouped into three subgroups: 3a, characterised by intraepithelial lymphocytes with mild villi atrophy; 3b, characterised by intraepithelial lymphocytes with marked villi atrophy; and 3c,

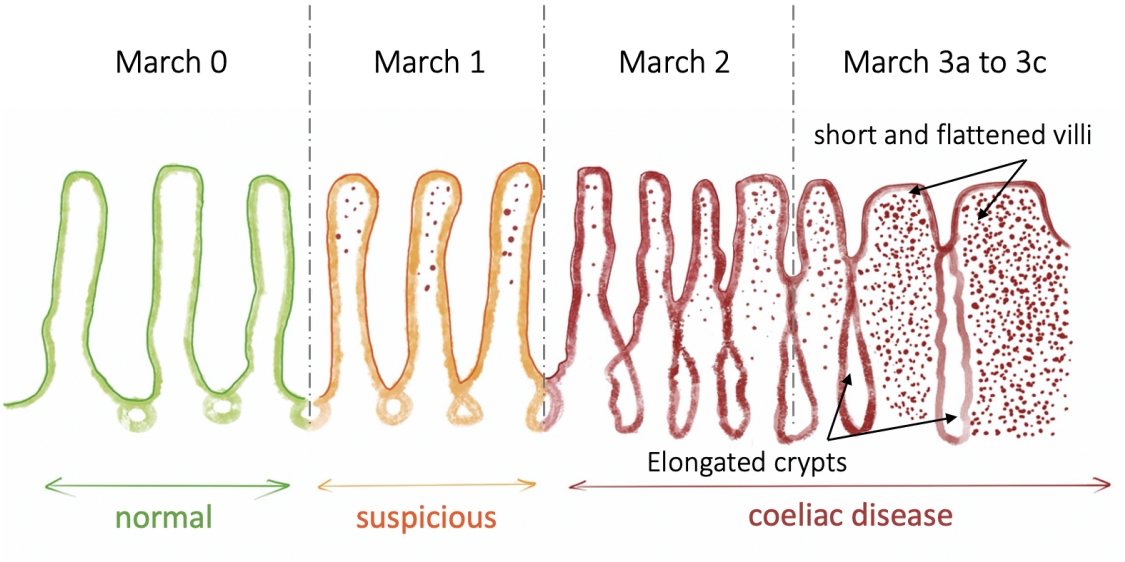


Figure 5.8: Schematic diagram of marsh classification in coeliac disease.

March Type	IELs	Crypt hyperplasia	Villi
0	<30	Normal	Normal
1	>30	Normal	Normal
2	>30	Increased	Normal
3a	>30	Increased	Mild atrophy
3b	>30	Increased	Marked atrophy
3c	>30	Increased	Complete atrophy

Table 5.1: Marsh classification of histological characteristics in coeliac disease

characterised by complete villous atrophy, and the mucosal profile resembles colonic mucosa. Table 5.1 presents Marsh classification of histological characteristics in CD.

Dataset Description. In collaboration with Dr. Holm Uhlig, we constructed this dataset by collecting whole-slide images of patients who underwent duodenal biopsy for the primary purpose of diagnosing coeliac disease from 2013 to 2016. All of these patients, aged between 5 months and 18 years, are part of the AbCD (Antibody Diagnosis in Paediatric Celiac Disease) cohort and have been consented to the Oxford Gastrointestinal Illness Biobank. The relevant diagnostic slides and the associated clinical data (e.g. age) were collected from Cellular Pathology Department

March Type	No. of cases	No. of slides
0	79	108
2	5	50
3a	17	102
3b	25	99
3c	7	33

Table 5.2: Statistics of the coeliac disease dataset

at the Oxford University Hospitals database. These slides were scanned by the Department of Surgical Sciences, Nuffield, University of Oxford using a Philips Ultra Fast Scanner running version 1.8.1.3 at 40x magnification. Consequently, we collected 2 to 7 slides from each patient, resulting in a total of 392 whole-slide images from 133 patients. The distribution of the cases and whole-slide images are depicted in Table 5.2. We randomly partitioned 295 of these whole-slide images for model training and 97 whole-slide images as a separate test set for the final evaluation of our model. The diagnosis of CD severity based on Marsh scores is given by pathologists as groundtruth for the classification task. To further estimate how the proposed model would generalise to the dataset, 2-fold cross-validations were performed. Each cross-validation model was trained using the same hyper-parameter and data preprocessing techniques.

Data Preprocessing. Typically, attributed to the extremely large size, whole-slide images have to be decoupled into numerous smaller patches that are independently processed by convolutional neural networks (CNNs) for training and inference. The routine tessellation strategy chops the images by a sliding- window moving across the entire whole slide leading to a large proportion of non-representative patches. In addition, the distinguishing features for CD severity diagnosis can be found in distinct regions on the whole-slide images that can be picked up by a preprocessing step. Therefore, in this study, a simple and efficient preprocessing pipeline using k-means clustering and autoencoders is used to eliminate the less representative patches so that the proposed patch-level classification network can be trained more effectively and efficiently. The method consists of a convolutional autoencoder to

learn the embedded features of each patch and a following k-means algorithm for patch clustering. In particular, the whole-slide images were split into 128 x 128 pixel tiles. This size was chosen because it was found to be on a similar scale to the width of a villous extrusion. This means that the villi patches can be identified as a key feature and separated from the rest of the image as a region of interest. A feature vector was generated for each patch using a convolutional autoencoder, and the patches were clustered into 8 categories using the k-means algorithm. As a result, the number of samples that fall into the background clustering category accounts for about 50% of the samples, while the other categories (villous, crypt, stoma, etc.) are more evenly distributed. The centres of each patches belonging to the villous cluster were extracted to generate weakly annotated data points. As a result, a large number of 1024 x 1024 patches around the annotation points were cut, and sample patches are shown in Figure 5.9. A detailed comparison of this pipeline with other methods that focus on detecting discriminative patches on whole-slide images is beyond the scope of this paper, and I will address this issue in future work.

5.4 Experiments and Results

5.4.1 Implementation details

The proposed model is trained end-to-end under the same experimental settings in 4.4.1. The hyper-parameters: λ in Eq. 5.5 is set to be 1 in our experiments, and was reduced by 0.1 after every 50 epochs. All the trainable weights are initialised with the weights pertained on ImageNet. The networks are implemented with MXNet and trained on two NVIDIA GeForce 1080 Ti GPUs.

5.4.2 Evaluation and Analysis on Clinical Fungi Dataset

Quantitative comparison with competing methods This section evaluated the effectiveness of the proposed method by comparing it with several state-of-the-art fine-grained classification methods. All of the compared methods were trained with the same backbone network and computing environment. Although our earlier work has already made a significant improvement (2.9% relative accuracy gain),

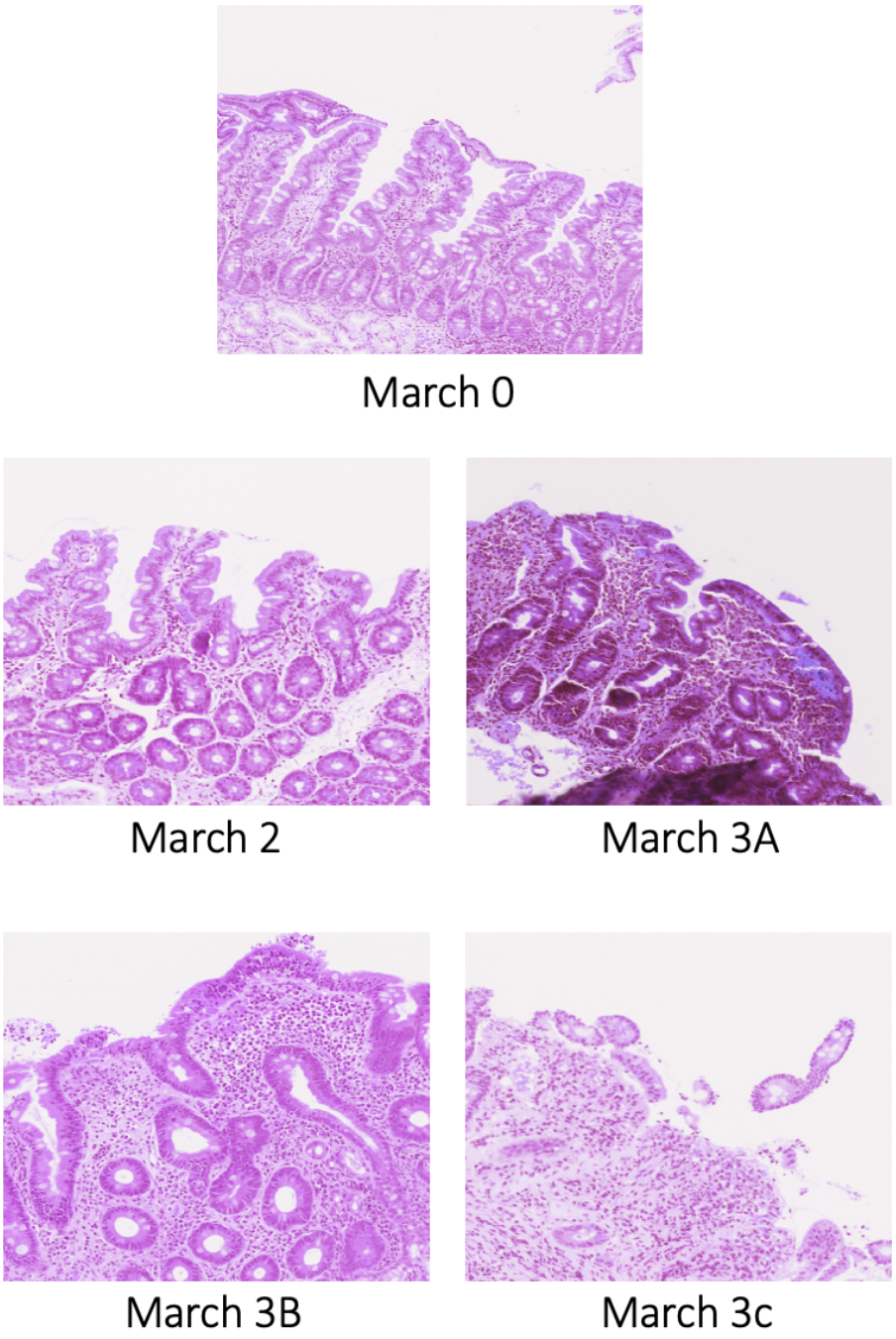


Figure 5.9: Marsh classification in coeliac disease. The histology images are selected from our coeliac disease dataset. March 0: normal; March 2: increased intraepithelial lymphocytes with crypt hyperplasia; March 3a: mild villous atrophy; March 3b: marked villous atrophy; March 3c: complete villous atrophy.

Methods	Accuracy
ResNet-50 [71]	0.907
RAN [113]	0.867
ARN [112]	0.871
TAN [119]	0.883
NTS Network [153]	0.914
Our earlier work [175]	0.943
Our method	**0.959**

Table 5.3: Experimental results on Fungal dataset.

it is computationally expensive and handles all sampled instances independently. From the comparisons shown in Table 5.3, we can observe that the proposed method achieves the best performance compared to other fine-grained classification methods, while showing an improvement of 1.6% compared to the method in the previous chapter.

Comparative evaluation of gated attention mechanism. To measure the effectiveness of the proposed gated attention module, experiments over various attention mechanisms were conducted. The same sampling strategy, feature fusion scheme and loss function were used in all experiments when comparing the attention mechanisms. Specifically, I only modified the attention-based instance localisation module in the baseline work, and investigate the performance of different attention mechanisms. Table 5.4 shows the results of integrating different attention modules in the proposed classification framework. It can be observed that our gated attention mechanism outperforms all other attention modules, demonstrating its effectiveness and localisation capabilities. Also, our lightweight gated mechanism is more efficient and has no additional computational overhead in the attention module.

Ablation study of effects of enhancement modules. This work extends the previous chapter with three main enhancements. In order to verify the effectiveness of each extension in this work, ablation studies with and without a gated attention module, a weighted feature fusion scheme, and an instance representation loss were conducted. The experimental results for different settings are shown in Table 5.5.

Attention Mechanisms	Accuracy
Spatial attention [117]	0.859
Channel-wise attention [119]	0.883
Dual attention [110]	0.901
Squeeze-Excitation attention [114]	0.939
Global Context attention [120]	0.937
Our earlier work [175]	0.943
Our Gated Attention	**0.951**

Table 5.4: Comparison of global attention mechanisms.

The baseline model is an early implementation showed in the previous chapter. It can be observed that all three enhancements push the classification performance further away from the baseline, confirming the contribution of these components in the proposed classification framework. Specifically, the gated attention module and the weighted feature fusion scheme achieve a relative improvement of 0.8% and 0.7%, respectively, while the instance representation loss achieves a relative accuracy improvement of 0.4%. This section also investigated the performance of integrating the two extensions together and made further improvements, which validated the compatibility of our proposed extensions. Finally, the combination of the three enhancements improved the classification accuracy to 0.959, improving the baseline framework by 1.6%.

In practice, the number of parameters and the average runtime for a single image are reported in Table 5.6. For all experiments, runtimes were measured on two Nvidia Geforce GTX 1080 GPUs. Compared to a standard single-stage network, the proposed network consists of two sub-networks, which is not sufficient to achieve real-time performance. However, the average running time for a single fungal image in this network is reduced by 0.02 seconds compared to the original design described in Chapter 4. This is mainly attributed to the design of the gated mechanism. In the original design, a hard filter was applied to explicitly select the first 100 channels, for which we had to first convert the tensor objects into NumPy arrays in Python. In contrast, the gated layers proposed here can be computed entirely based on tensor operations and can therefore be accelerated using the GPU.

Gated attention	Weighted fusion	Instance loss	Accuracy
			0.943
✓			0.951
	✓		0.950
		✓	0.947
✓	✓		0.954
✓		✓	0.953
	✓	✓	0.951
✓	✓	✓	**0.959**

Table 5.5: Ablation study of our method on fungi dataset

Method	Number of parameters	Run time
ResNet-50	25,549,480	0.02s
Original design	49,562,198	0.06s
This work	49,037,910	0.04s

Table 5.6: Average runtime for a single image in the clinical fungal dataset. The original design refers to the network described in chapter 4, while this work is the architecture presented in this chapter.

Qualitative visual evaluation of gated mechanism A visual example of each step in the gated attention module is depicted in the Figure. 5.10, which shows that the confounding channels can be discarded and only meaningful channels are retained. Given an input image, the first row displays the visualisation of the feature maps in the last convolutional layer and the generated global semantic map, while the second row presents the final attention map and two representative channels. In particular, the attention weight is approximately 1 for the most informative channels, and approximately 0 for the channels highlighting the irrelevant background. Consequently, our gated attention mechanism not only outperforms all other attention modules but it also suppresses the confounding information, demonstrating the effectiveness and localisation capabilities.

Qualitative visual evaluation of instance localisation Following up on the failure cases shown in Figure 4.13, this section visualises the corresponding attention maps generated in the network presented in this chapter, as shown in Figure 5.11. These five images in the first column were misidentified by our early network due to large variations in illumination and structure, but were successfully classified in our

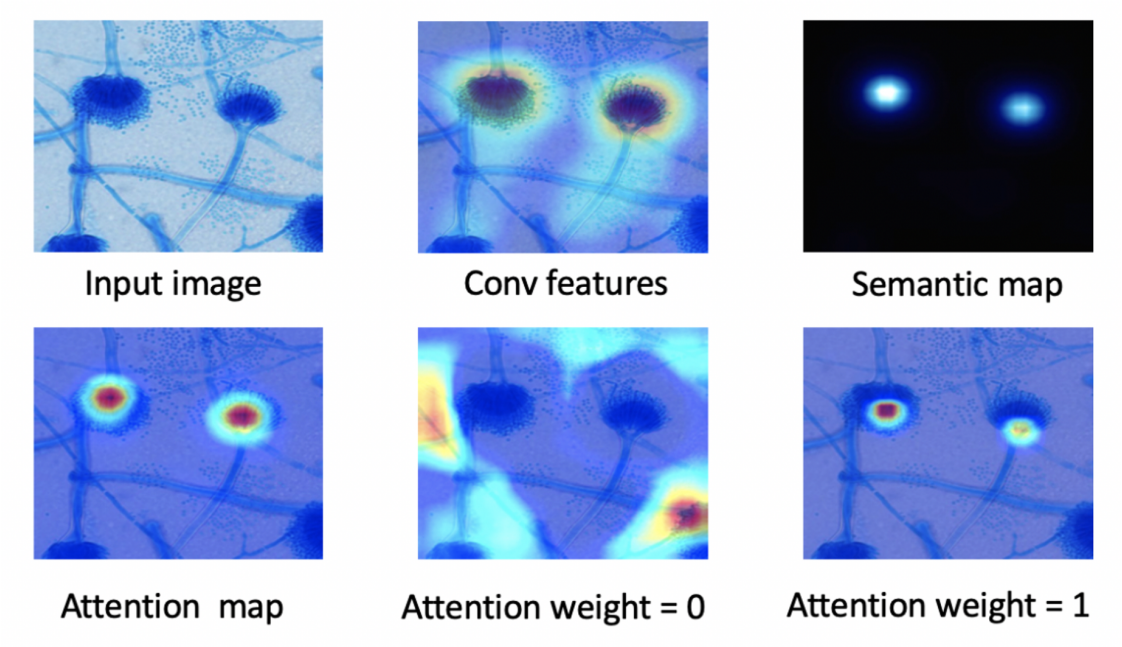


Figure 5.10: Visualisation of the gated attention module for one sample image. The top row shows the input image, original convolutional feature map and the corresponding semantic map (defined in Eq.5.1). The bottom row shows the final attention map and two representative channels with the lowest and highest attention.

proposed work. It is shown that our new extensions enhance the robustness of the network to different image conditions and can effectively localise the discriminative instances under challenging conditions, while discarding irrelevant background information for subsequent instance-level feature learning.

In addition to the successful results, three interesting cases of fungal images where the proposed method failed are shown in Figure 5.12. It can be seen that the proposed gated attention module has difficulty detecting the exact vesicle region when there are very large instances (second image in Figure 5.12) or clusters of small instances (first and third images in Figure 5.12). For example, as in the second image in Figure 5.12, the network tends to extract two separate instances from this large fungal object, which may prevent the following instance-level network from extracting information about the shape of the vesicles.

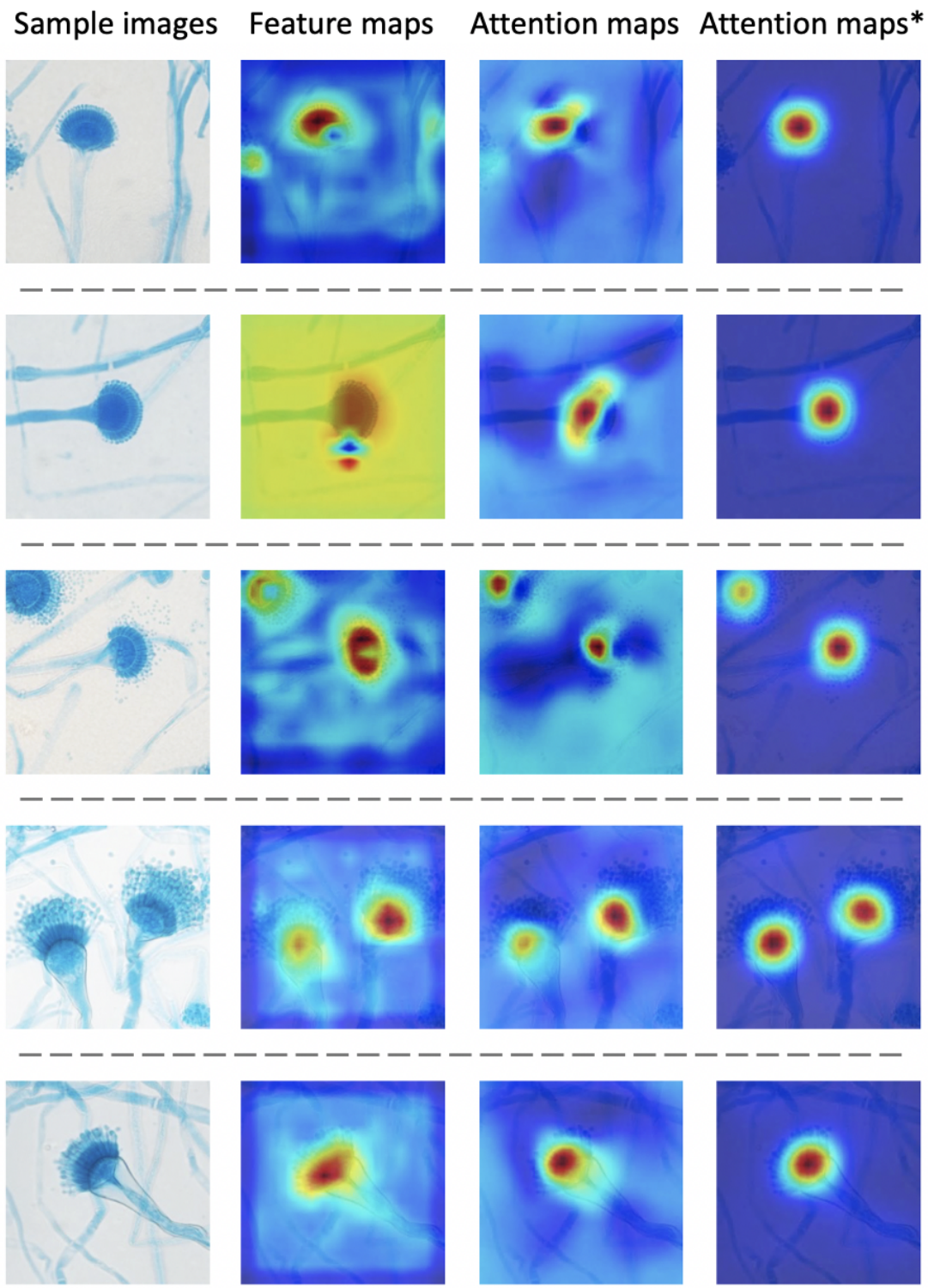


Figure 5.11: Visualisations of the attention mechanism in fungi dataset. This figure shows five challenging cases, which were misclassified by our earlier baseline method, but were correctly addressed in this network. The attention maps in the third column were generated by the network described in Chapter 4, and the attention maps in the last column were produced by architecture described in this chapter.

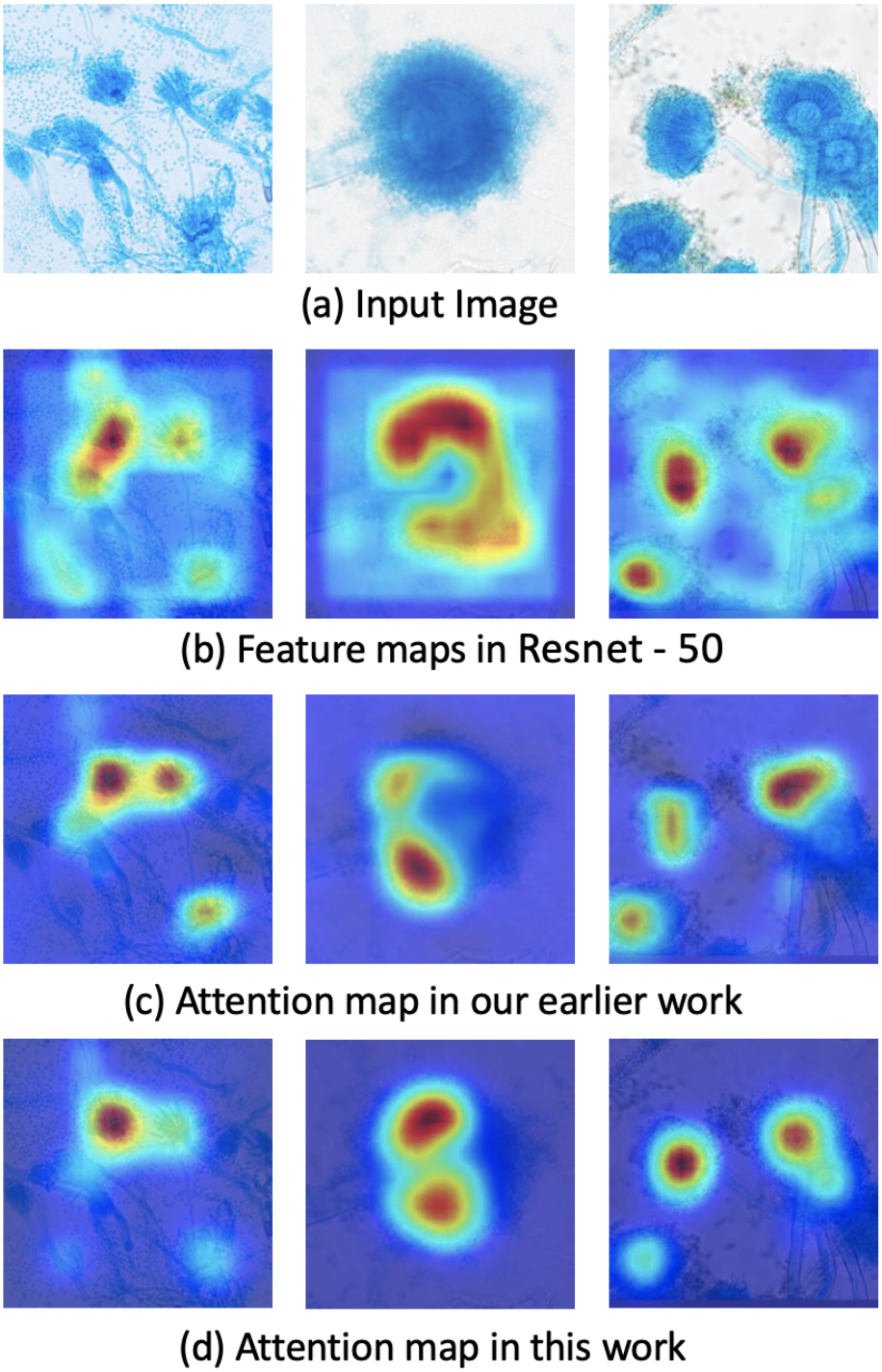


Figure 5.12: Visualisations of the failure cases within the test set. It can be seen that when there are very large instances or clusters of small instances, it is difficult for the model to detect the exact object boundaries.

5.4.3 Evaluation and Analysis on BACH dataset

Table 5.7 shows the classification results on breast cancer histology images. I compared the best classification accuracy of several advanced methods in the case of 400 images provided by the challenge organiser. Kohl et al. [176] investigated the applicability of general CNN-based networks to the problems of histology image classification, and showed that DenseNet-161 network achieved the classification accuracy of 0.940 that outperformed other CNN models. Rakhlin et al. [82] fused the deep feature descriptors extracted by ResNet-50, VGG-16 and Inception-v3 networks, and adopted gradient boosted trees for the final classification. Nawaz et al. [177] fine-tuned an Alexnet by changing and inserting convolutional layers. Cao et al. [72] presented a deep learning-based Random Forest kernel SVM (RFSVM) and achieved the highest accuracy of 0.930. Although promising performance has been achieved, these methods can hardly unitise the high-resolution information of medical images due to the downscaling operation. Some recent works [73–75] also introduced two-stage frameworks that first extracted smaller high-resolution patches and conducted patch-wise classification. The final image-wise labels were obtained by combining the classification results of all extracted local patches. However, the pre-extracted patches may ignore the long-range correspondences and suppress the important fine-grained information, such as the density or shape of objects. Besides, how to aggregate the patch-level information in a meaningful way still needs to be investigated carefully. In contrast to these methods, the proposed method integrates local instances and global structural features, where relevant high-resolution information can be obtained by resampling discriminant patches. As a result, our approach achieves the best classification performance with 0.970, showing that our network can be effectively applied to the classification tasks of histology images.

5.4.4 Evaluation and analysis on ISIC-2017 Dataset

Melanoma is the most dangerous type of skin cancer that has the highest mortality rates. Recently, a number of deep learning-based melanoma recognition methods

Methods	Accuracy
Vgg19 [176]	0.925
Inception-v3 [176]	0.913
DenseNet-161 [176]	0.940
Model Fusion [82]	0.925
AlexNet [177]	0.813
ResNet-152 [72]	0.830
RFSVM-All [72]	0.930
Ensemble [73]	0.825
Ensemble with refinement [73]	0.875
Two-stage network [74]	0.850
Hybrid deep network [75]	0.913
Our earlier work [175]	0.960
Our method	**0.970**

Table 5.7: Experimental results on breast cancer dataset.

Methods	Ext. Data	Sen	Spe	Acc	AUC
#1[92]	7544	0.547	0.950	0.873	0.874
#2[88]	1600	0.427	0.963	0.858	0.870
#3[86]	1444	0.735	0.851	0.828	0.868
#4[99]	900	0.103	0.998	0.823	0.856
#5 [89]	1341	0.350	0.965	0.845	0.836
#6[178]	0	0.436	0.925	0.830	0.830
#7[179]	0	0.521	0.901	0.827	0.805
#8[180]	0	0.376	0.957	0.843	0.804
RAPooling [76]	0	0.607	0.884	0.830	0.842
Ensembles [87]	0	0.402	0.719	0.852	0.851
ResNet*[77]	0	0.529	0.905	0.832	0.789
Attention* [78]	0	0.590	0.896	0.837	0.859
This work	0	0.522	0.938	0.851	0.861

Table 5.8: Experimental Results on ISIC-skin 2017 test data.

have been proposed. Table 5.8 gives the performance in melanoma classification on ISIC-skin 2017 official test data. I compared our proposed network with the top 8 methods in the challenge leaderboard and four other methods proposed recently. The classification performance was assessed by sensitivity (Sen), specificity (Spe), accuracy (Acc) and area under the receiver operating characteristic curve (AUC). In the ISIC 2017 skin lesion challenge, the AUC score was used as the golden evaluation index by which all participants were ranked. Other evaluation

metrics are defined as follows:

$$Sensitivity\,(Sen) = \frac{TP}{TP + FN} \tag{5.6}$$

$$Specificity\,(Spe) = \frac{TN}{TN + FP} \tag{5.7}$$

$$Accuracy\,(Acc) = \frac{TP + TN}{TP + FP + TN + FN} \tag{5.8}$$

where TP, FP, TN and FN represent the number of true positives, false positives, true negatives and false negatives.

As shown in Table 5.8, the top five methods listed in the leaderboard all use additional dermoscopic images to expand the training set and boost their classification performance. In particular, the top-ranked solution [92] uses as many as 7,544 external images. However, our network was trained only on ISIC-2017 training data, yet achieved comparable performance in melanoma classification. Based on the ranking rule of the challenge, our method achieved an AUC of 0.861, higher than #4 [99] and #5 [89] which both used additional training data.

Furthermore, I compared the network with seven methods using only the official training set. Yang et al. [76] proposed an end-to-end melanoma recognition framework that used the segmented lesion masks to guide the classification performance through a new pooling operation named region average pooling. The proposed framework consists of a regular ResNet-50, followed by two branches: segmentation branch and classification branch. Harang et al. [87] proposed an ensemble-based approach for skin lesion classification. They pre-trained four different CNN architectures, and the final prediction was achieved based on the weighted output of the member CNNs. Although they achieved the highest accuracy among these methods, this approach had extremely high computation cost and lowest specificity. Besides, a CNN-based regularised discriminative learning framework was proposed in [77], which minimised the whole of within-class variance and maximised the inter-class variance. Zhang et al. [78] incorporated a novel attention mechanism to enhance the discriminative representation ability of the deep network. From the comparison in Table 5.8, we can see that the proposed network achieved the

highest AUC and the second highest accuracy without the use of external images, lesion segmentation and ensemble strategies.

5.4.5 Evaluation and Analysis on Coeliac Disease Dataset

The baseline classification model was a standard ResNet-50 [71] trained by the entire training dataset with 35467 patches. I find that the classification accuracy produced by ResNet at the patch level is 76.6%, while the average accuracy improves to 84.8% when using the proposed attention-based multi-instance framework. The effectiveness of generated attention maps was assessed by visualising models' predictions to highlight the most informative regions relevant to the predictions (Figure 5.13). The first column is sample images from the coeliac disease dataset, and the second column is the visualisation result of the last convolutional layer in the ResNet [71], while the last column shows the corresponding attention maps generated in our network. As mentioned previously, clinical diagnosis is based primarily on the analysis of villous atrophy, crypt hyperplasia and intraepithelial lymphocytosis. The attention maps generated by our network indicate that our model has learned the correct clinically relevant features where informative regions can be effectively detected. In addition, I evaluated the performance of the proposed method for classifying coeliac disease into five different grades through 2-fold cross-validation. In the whole-slide inference, the most commonly predicted class at the patch level was chosen as the whole-slide prediction. Table 5.9 and Figure 5.14 depict the classification results for each class. In general, the proposed method achieved high performance in the detection of normal cases, while the classification accuracy of Marsh 3c cases was the lowest across the board. There are two reasons for this. One reason could be that Marsh 3c cases had the lowest number of training samples, with only 33 whole-slide images in the dataset, compared to 108 and 102 for normal and Marsh 3a cases, respectively. Another possible reason, which has been confirmed by pathologistsis, is that this category is characterised by complete villous atrophy, i.e., the apparent absence of villous structures, and since we first used a cluster-based pre-processing pipeline to extract villous patches, it is likely

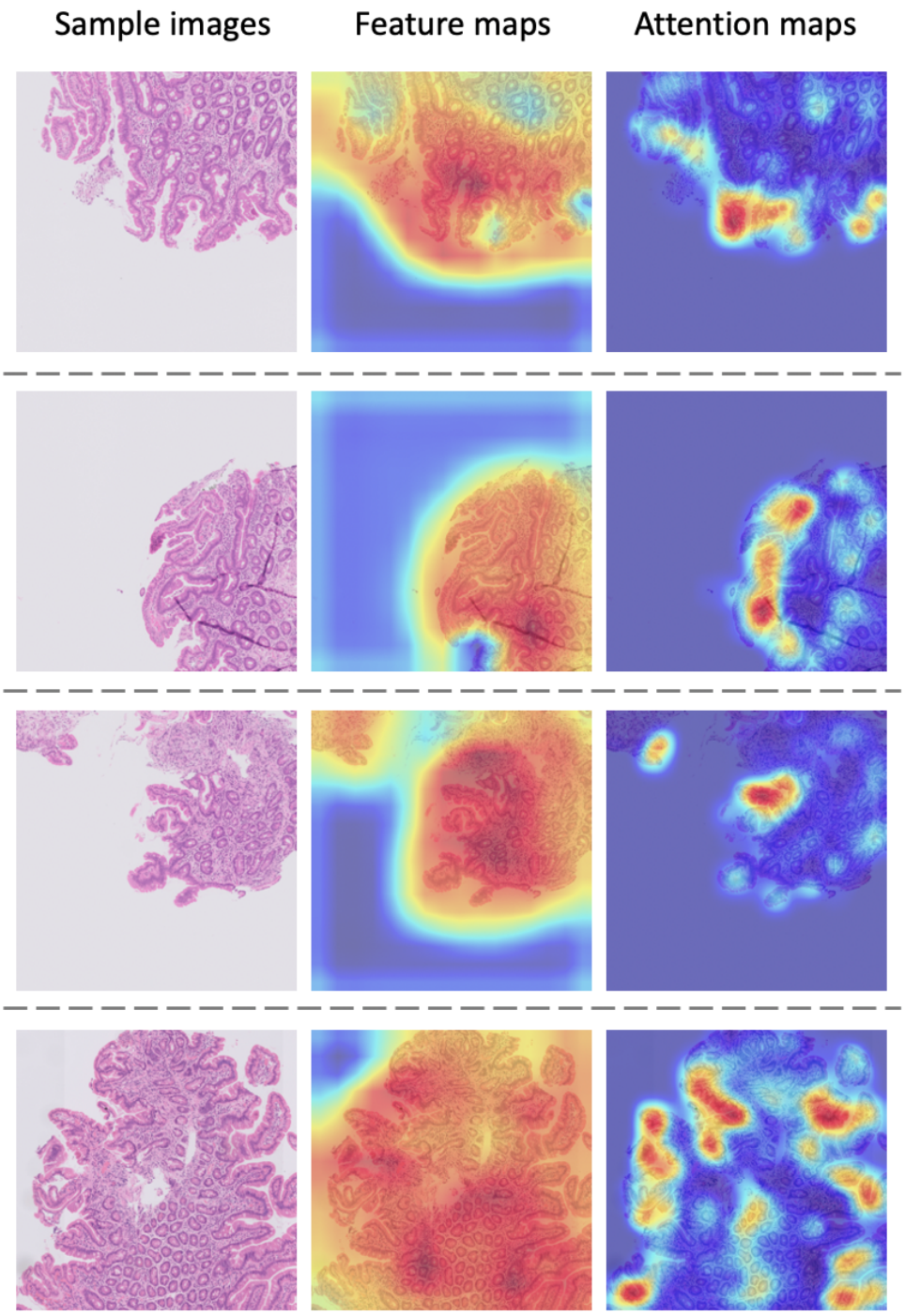


Figure 5.13: Visualisation of the attention maps for coeliac disease dataset. The first column shows sample images from the coeliac disease dataset, and the second and third columns visualise the results of the last convolutional layer in the ResNet and the attention maps generated in our network, respectively. It can be observed that ResNet tends to focus on the entire tissue regions, whereas our network can explicitly identify the informative villous regions.

March Type	Fold 1		Fold 2	
	Patch	Image	Patch	Image
0	0.974	1.000	0.973	1.000
2	0.847	0.916	0.863	0.916
3a	0.820	0.833	0.813	0.802
3b	0.794	0.840	0.809	0.852
3c	0.647	0.500	0.641	0.500

Table 5.9: Experimental results on coeliac disease dataset.

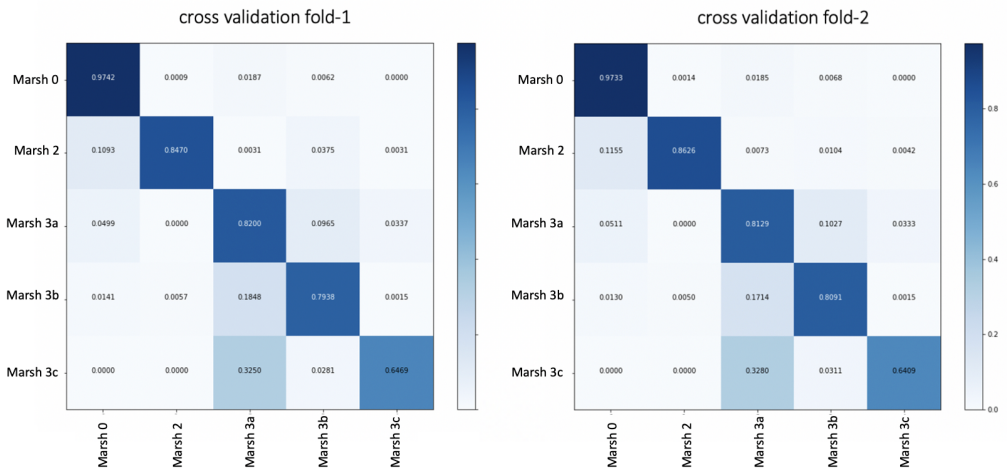


Figure 5.14: Confusion matrices showing the classification performance on coeliac disease dataset.

that some non-informative patches were also extracted in the patch-level dataset due to the difficulty in locating villous regions in Marsh 3c cases, making it more challenging to detect this category. It is important to note the core purpose of these experiments was to test the ability of the proposed architecture to estimate the severity of coeliac disease at the patch level. Future research will compare it with more state-of-the-art classification methods and explore more flexible aggregation algorithms for whole-slide inference.

5.5 Clinical Background Alignment

In recent years, vision systems based on deep learning have progressed by leaps and bounds, rivalling the performance of human experts even in applications that require domain knowledge, such as biomedical image analysis. However, as these AI-powered vision systems become more sophisticated, they also become more complex and

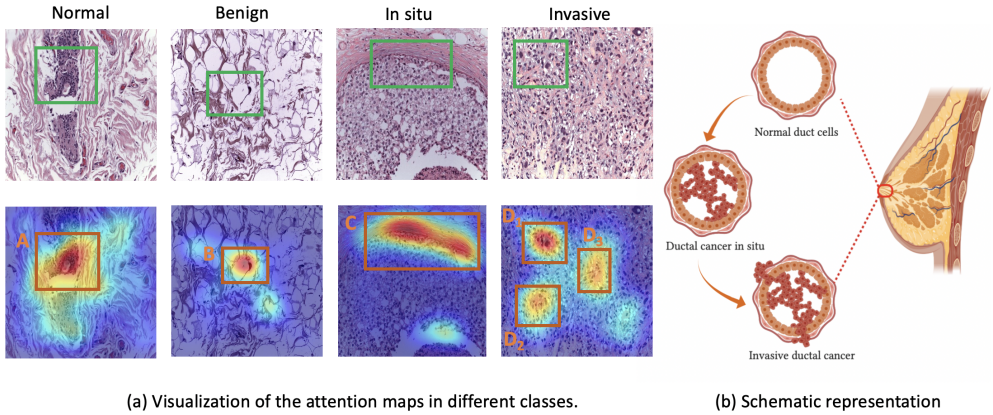


Figure 5.15: Clinical alignment on breast cancer dataset. The first row in (a) shows the sample images with bounding box annotations generated by a pathologist. The corresponding attention maps shown in the second row consistently focus on the clinically relevant features. The localised discriminative region B corresponds to the loose connective tissue. In carcinomas, region C highlights the boundary of the ducts, while D_1, D_2 and D_3 focus on the poorly organised clusters of cells, in which cancerous cells spread into surrounding tissues. (Figure (b) was created with `https://BioRender.com`.) Notice that the bounding boxes on the generated attention maps were placed manually by myself.

opaque. The reader can easily appreciate the importance of transparency in decision making and interpretability of learned features (alignment with medical knowledge) for the trustworthiness of such automated systems in biomedical scenarios. This section delves into the details of interpretability of our proposed method that can be used as a general explainable approach for biomedical image classification problems.

5.5.1 Breast Cancer Classification from Histopathology Images

In clinical practice, pathologists routinely make histopathological assessments by inspecting certain morphological changes under the optical microscope, which is considered as the gold standard for diagnosis of breast cancer. While examining the breast tissue samples, pathologists typically assess the tissue structure patterns and morphological characteristics of different components across the tissue. The schematic view of the progression from normal breast tissue to invasive breast cancer is illustrated in Figure 5.15 (b). Recently, due to the dramatic improvement of image techniques, there is an increase in the availability of computer-aided diagnostic

systems. Figure 5.15 (a) shows the examples of different types of breast tissue, and the corresponding attention maps generated from our network. It can be readily observed that in addition to our method being able to correctly identify the types of breast cancer, most attention maps are also remarkably consistent with the clinical-pathological features, demonstrating our ability to understand semantic biomedical patterns, which is a significant contribution to explainability and hence trustworthiness of AI-based vision systems in biomedical applications.

To qualitatively evaluate the effectiveness of learned features, bounding box annotations were generated by an expert pathologist on 100 test images. For each bounding box annotated by an expert, the annotated boundaries are detected by the proposed network if most of its area overlaps with the boundaries I placed myself. Otherwise, this annotated bounding box is missed when the majority of its pixels did not match any of the bounding boxes generated on the attention map. In total, 72% of the bounding boxes were covered by the network based on visual comparisons. Specifically, a normal healthy breast duct is made up of layers of inner epithelial cells, outer myoepithelial cells and a basement membrane. Our network explicitly highlights the normal cuboidal epithelial cells (region A in Figure 5.15 (a)), which are well organised in the regular duct. Benign diseases, though usually not harmful, but present greater morphological complexity. For these benign lesions, pathologists mainly search for signs of stromal changes around the duct. Similar to the features observed in clinical assessment, the localised discriminative region B corresponds to the loose connective tissue. In carcinomas, the surrounding micro-environment and normal tissues fail to function properly, resulting in dramatic changes in cell growth and tissue organisation. Like the typical patterns shown in Figure 5.15 (b), in the cases of *in situ* carcinomas, the proliferating cancer cells are restrained inside the duct and the basement membrane remains intact. Later, as the myoepithelial cells and the basement membrane are disrupted, the cancer cells break out of the walls of ducts and invade the surrounding breast tissues, leading to invasive carcinomas. As a result, the intactness of the basement membrane can be used as an important factor to differentiate *in situ* and invasive breast cancers.

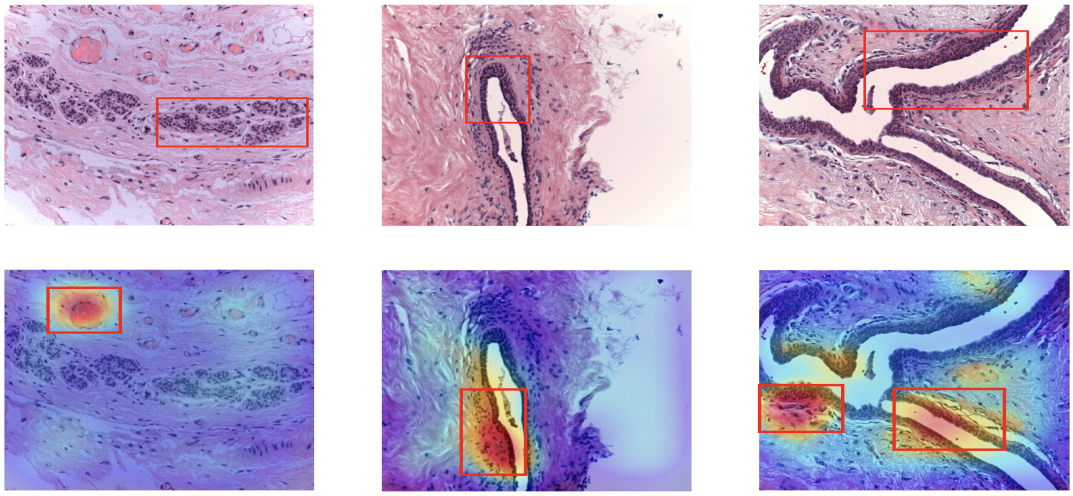


Figure 5.16: Alignment errors in the generated attention maps and expert annotations.

In Figure 5.15 (a), region C highlights the boundary of the ducts, while D_1, D_2 and D_3 focus on the poorly organised clusters of cells, in which cancerous cells spread into surrounding tissues. In contrast, Figure 5.16 describes examples where network attention failed to explicitly target those bounding boxes annotated by the pathologist. First, it is known that there is some inter- and intra-observer variation in the annotation of biomedical images, which means that different pathologists may look at different areas to make their decisions. All of the annotated bounding boxes used in this study were collected from one pathologist, and therefore the salient regions highlighted by the proposed network will likely contain histopathological patterns that are meaningful to other pathologists. In addition, due to the limited size of this dataset, the network may be sensitive to the colour, illumination and some background patterns in the dataset, as in the first image in Fig. 5.16.

5.5.2 Skin Cancer Classification from Dermatoscopic Images

In clinical practice, the current guidelines for melanoma detection are mainly based on the ABCD rules (Asymmetry, Border irregularity, Colour variegation, and Diameter), as described in Figure 5.17 (b). The ABCD rule [181] provides the dermatologists with a comprehensive template of clinical features of the malignant

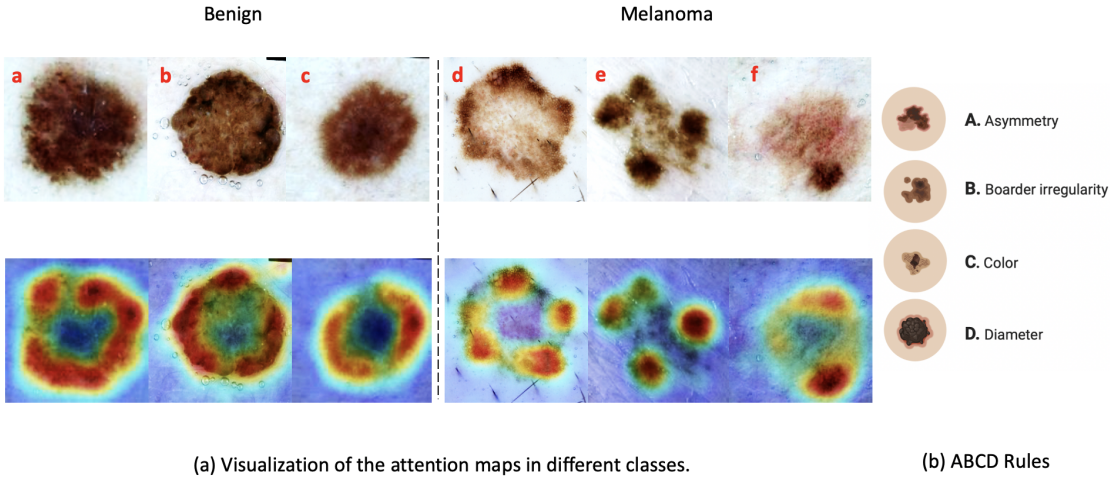


Figure 5.17: Clinical alignment on ISIC-2017 skin cancer dataset. The attention maps in benign cases highlight the regular round boundaries, whereas in melanoma cases the network tends to focus on the areas with major irregularities in colour space. (Figure (b) was created with `https://BioRender.com`.)

melanoma: (1) the shape of melanoma is usually asymmetrical; (2) the boundaries are irregular and poorly defined; (3) the colour is varied among the lesion area; and (4) the diameter is usually greater than 6mm. Several examples of benign and malignant melanoma are presented in Figure 5.17 (a) for comparison.

The visualised attention maps provide us with some insight into whether the network makes the decision based on the clinically relevant regions. For benign cases, the regular round boundaries are highly activated. In particular, the third one (c) highlights the borders on the opposite side of the lesion, explicitly indicating the symmetry in shape. On the other hand, for melanoma cases, the network usually focuses on the areas with major irregularities in colour space. However, the diameter criterion is less effective in our experiments since over 50% melanoma lesions are smaller than 6mm in this dataset.

5.5.3 Fungal Species Classification from Mycology Images

Traditionally, clinical laboratories rely primarily on morphology-based methods to differentiate Aspergillus at the species level, such as A. flavus, A. fumigatusand,

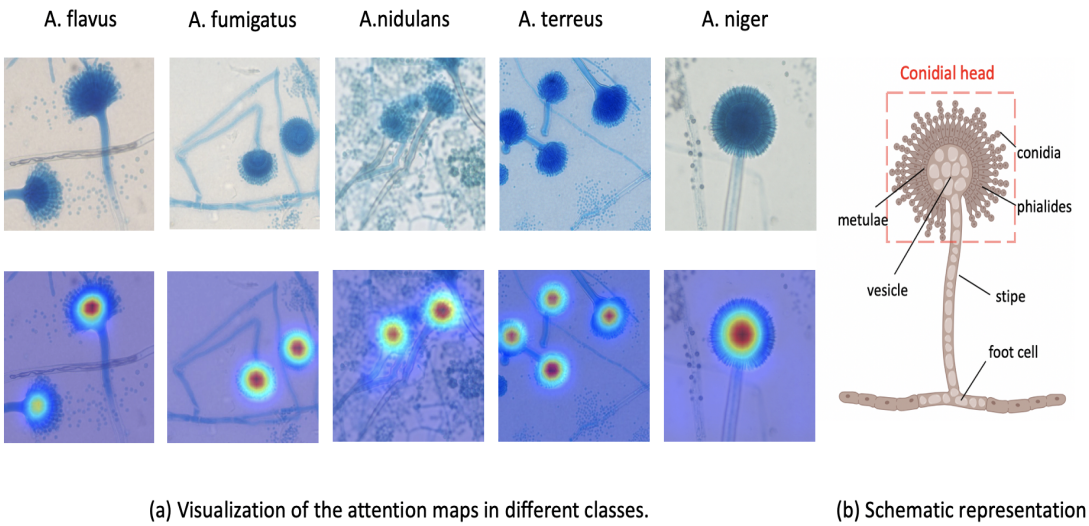


Figure 5.18: Clinical alignment on fungi dataset. The attention maps of different species consistently focus on the regions of conidial heads, especially the regions of vesicles. (Figure (b) was created with `https://BioRender.com`.)

A. nidulans, A. terreus and A. niger. The principal morphological characteristics of different species were outlined in Section 4.3. Generally, the key criteria in species identification are a range of morphological features associated with conidial heads which consist of the vesicles, metulae (if present), phialides and conidia (Figure 5.18 (b)). Specifically, the vesicles from different species usually vary in size and shape, which is usually considered as the most important cue for Aspergillus identification. Based on these criteria, the visual analysis was conducted to measure the alignment of our method with the clinical background. Figure 5.18 (a) shows the sample images and corresponding attention maps of each species. All attention maps consistently focus on the conidial heads, especially the regions of vesicles. Therefore, it can be seen that the discriminative regions outlined by our network effectively match the areas mycologists look at. It is, however, worth noting that our method also highlights the foot cells in A. nidulans cases. Although the stipe structure is not an important factor for Aspergillus identification, the stipe of A. nidulans is usually the shortest among the five species in our dataset, leading to the existence of foot cells. Hence, in our experiments, the region of foot cells are

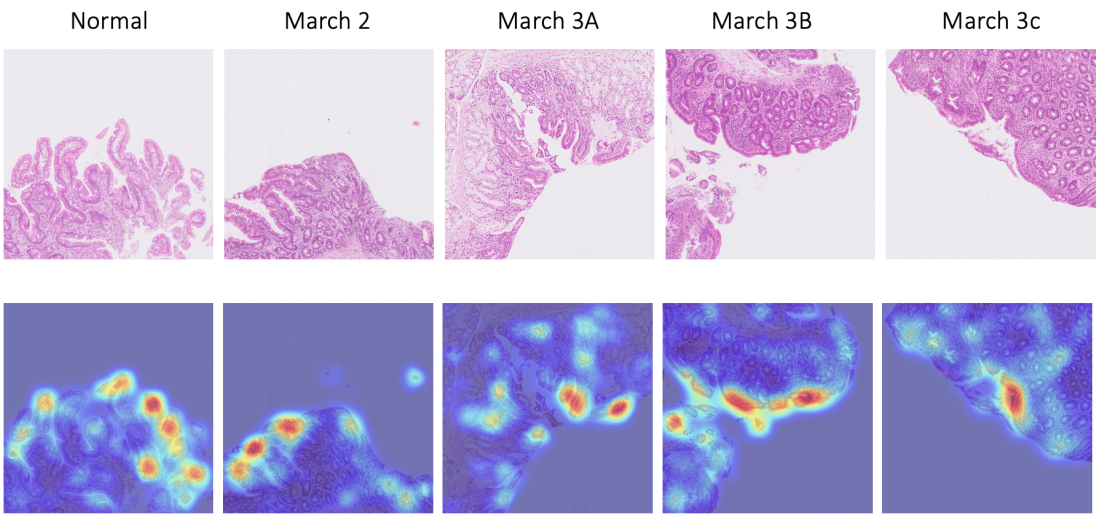


Figure 5.19: Clinical alignment on coeliac disease dataset. The attention maps consistently highlight the most clearly defined villous structures associated with different degrees of severity: normal or mild samples show a preserved villous structure whereas severe cases highlight the shortened and flattened villi.

also important in distinguishing different Aspergilus species. This is an important insight into the interpretability of the proposed network and the explainability of the automated decision process.

5.5.4 Coeliac Disease Classification from Histopathological Images

As described in Section 5.3.4, the clinical diagnosis of coeliac disease severity is primarily based on the histological features observed in duodenal biopsies, including (1) increased intraepithelial lymphocytes (IELs), (2) crypt hyperplasia, and (3) partial to complete atrophy of the villi (Figure 5.8). I directly compare the attention maps generated in the proposed network with the reference standards established in the Marsh grading system. The visualisation results of the selected attention regions (Figure 5.19 and Figure 5.20) indicate that our model has the capability to localise regions of interest that meet clinical diagnostic criteria. Specifically, the attention maps in Figure 5.19 consistently highlight the most clearly defined villous structures associated with different degrees of severity: normal or mild samples show a preserved villous structure without atrophy or chronic inflammation, whereas severe cases of

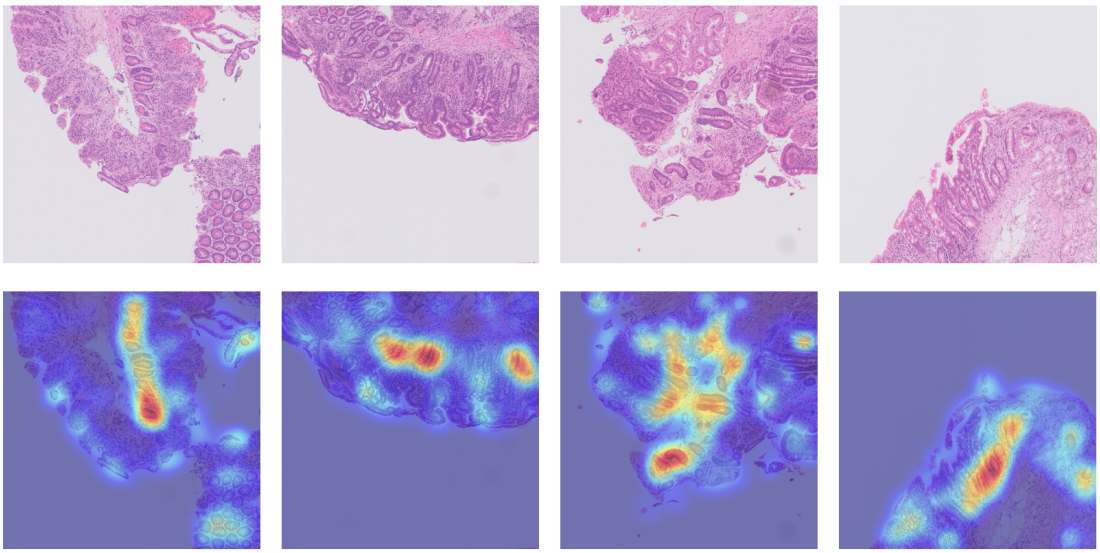


Figure 5.20: Visualisation of attention maps highlighting crypt hyperplasia in disease samples.

coeliac disease highlight histological structural changes, i.e. shortened and flattened villi. In terms of the diseased samples, Figure 5.20 particularly visualises some attention maps highlighting the regions of crypt hyperplasia (elongated crypts)

In summary, this approach works well for problems with relatively simple and consistent clinical criteria, such as fungal species identification. Although the generated attention maps do not always capture the clinical complexity of the disease, qualitative results have shown that deep learning-based approaches have great potential to assist pathologists in identifying regions of interest and providing initial diagnoses in the clinical workflow.

5.6 Conclusion

This chapter presented three technical extensions to the work presented in Chapter 4. It was achieved by first proposing a lightweight gated attention mechanism that can detect multiple critical instances simultaneously, followed by a weighted feature fusion scheme and instance representation loss to achieve the final image-level prediction. The proposed network was evaluated in four different biomedical scenarios: fungal species classification, breast cancer classification, melanoma

detection, and coeliac disease severity diagnosis. The quantitative and qualitative results show that it is a powerful general approach that can be applied to a variety of such fine-grained, multi-instance biomedical imaging problems to achieve state-of-the-art classification performance. More importantly, this chapter presented in detail the scope of the proposed method as an interpretable model applicable to a wide range of similar biomedical image classification problems by showing strong alignments of the learned features with medical knowledge. We believe that the work presented here opens up a number of avenues for further whole-slide image-based research and paves the way for deep learning models to gain the trust of the medical community.

Delays in dataset collection and experimental pipeline prevented us from conducting extensive experiments on the coeliac disease dataset. Thus, there is still some room for improvement in extending the proposed framework to the whole-slide image classification. For instance, the patch-level classification results were simply aggregated based on the majority voting, which ignores the contribution of different patches to image-level labels. Some avenues for further investigation in whole-slide image analysis will be discussed in Section 6.2.

6 Conclusions and Future Work

Contents

6.1 Summary of Contributions 153
6.2 Directions for Future Work 156
6.2.1 Region-based whole-slide image annotation strategy . . 156
6.2.2 Weakly Supervised Whole-slide Image Classification . . 159
6.3 Summary 162

This thesis has investigated and evaluated a series of automatic image analysis methods to translate the high-resolution biomedical imaging data into meaningful and quantitative information, with limited availability of training data and weak annotations. This chapter summarises the major contributions of this thesis while discussing the limitations of the current work and potential directions for future research.

6.1 Summary of Contributions

This thesis mainly addresses two specific research questions: cell segmentation (Chapter 3) and biomedical image classification (Chapter 4, Chapter 5). The main contributions are outlined as follows.

- **Global probabilistic model for dense cell nucleus segmentation.**

Chapter 3 introduces a hybrid method armed with a U-Net and a global probabilistic model for precise segmentation of dense cell populations as a preliminary step in analysing small biomedical datasets. The proposed global optimal model explicitly enforces prior shape constraints and global cell interactions without any user interaction. Our experiments demonstrate that the proposed global optimal model achieves superior segmentation performance and effectively compensates for the potential segmentation failures obtained with conventional CNNs. Although our current work focuses on the segmentation of monolayer cells, it is clear that this approach can be extended to 3D images and also applied to more complex cell types.

- **Gate attention mechanism for simultaneous localisation of multiple discriminative instances.**

  Compared to natural images, biomedical images contain numerous instances of interest (biomedical objects) with varying sizes and densities, resulting in more complex structural information and greater intra-class variation. Chapter 5 presents a novel gated attention mechanism that is capable of simultaneously localising multiple discriminative instances and eliminating the confounding effect of irrelevant regions without extra part supervision or redundant region proposals. In order to measure the effectiveness of our gated attention module, it was compared to other existing attention mechanisms where the same sampling strategy, feature fusion scheme, and loss function were used in our experiments for a fair comparison. The quantitative and qualitative results show that our gated attention mechanism not only outperforms all other attention modules, but it also suppresses confusing information on complex backgrounds, which proves its effectiveness and localisation capability in biomedical image analysis.

- **Multi-task learning scheme for fine-grained feature learning.**

  In contrast to other state-of-the-art two-stage frameworks consisting of two independent stages, this thesis presents a novel deep network for biomedical

image classification that jointly learns local instance-level and global image-level features in an end-to-end manner. An instance representation loss was proposed to enhance the internal consistency between the most discriminating instance within an image and all other sampled instances, potentially minimising intra-class variance. Specifically, the multi-task learning scheme consists of global attention loss, classification loss and instance representation loss, where the weights of member modules are dynamically controlled to force the network to learn consistent fine-grained features. The proposed framework achieves state-of-the-art performance on four different biomedical image analysis tasks: fungi species recognition, breast cancer classification, melanoma detection and severity diagnosis of coeliac disease.

- **Vision system with clinical alignment and improved interpretability.**

  While advanced deep learning techniques have greatly improved our ability to develop accurate and robust models for classifying biomedical categories, these black-box methods are incapable of elaborating the reasons behind their decisions due to unexplained latent representations of deep neural networks. Therefore, the lack of interpretability has been widely considered to be one of the major barriers to the incorporation of deep learning techniques into clinical applications. This thesis provides an interpretable attention-based framework for automatic classification in biomedical scenarios. Chapter 5 showcases the interpretability of learned features by visualising and comparing them with biomarkers that human experts focus on in their decision making. The visualisation results on four different biomedical datasets indicate that our attention maps consistently highlight the most important clinically relevant features.

- **Two new clinical datasets for biomedical image classification.**

  This thesis describes the two new datasets for biomedical image classification tasks. Chapter 4 describes the clinical background and collection details

of the fungal species dataset, which is comprised of the five most common human disease-associated species, including Aspergillus fumigatus, Aspergillus flavus, Aspergillus niger, Aspergillus terreus and Aspergillus nidulans. To the best of our knowledge, the ongoing project is the first attempt to bring deep learning-based approaches to fungal species identification. Chapter 5 also introduces a new whole-slice histological dataset for the diagnosis of coeliac disease severity. The biopsy image data from our project were scored by a professional pathologist, and the images were divided into five categories: normal, Marsh II, March IIIa, March IIIb, and March IIIc.

6.2 Directions for Future Work

6.2.1 Region-based whole-slide image annotation strategy

Chapter 4 and 5 presented automatic methods for localising multiple discriminative regions on high-resolution images (e.g. 1024 x 1024 pixels). The pipeline took advantage of global features extracted on downscaled images and used an attention mechanism to detect the discriminative instances on high-resolution images. While the classification framework has achieved good performance on four biomedical datasets, its performance and robustness may be further improved to the extent that it can be applied to large whole-slice histopathology images (typically 100,000 $\times$ 100,000 pixels). Owing to the extremely large scale and high resolution, it is difficult to extract meaningful information (e.g. small regions of interest at high resolution) directly from the downsampled version of the whole-slide images. Therefore, the whole-slide images have to be decoupled into numerous smaller patches that are processed independently by the deep networks for training and inference. The conventional tessellation strategy of slicing the whole-slide images by moving a sliding- window over the entire images results in a large percentage of non-representative patches. More importantly, the discriminative information of whole-slide images is prone to be suppressed when the information patches only occupy a small proportion of the cropped patches compared to the irrelevant patches. This results in the need for developing techniques to eliminate less representative

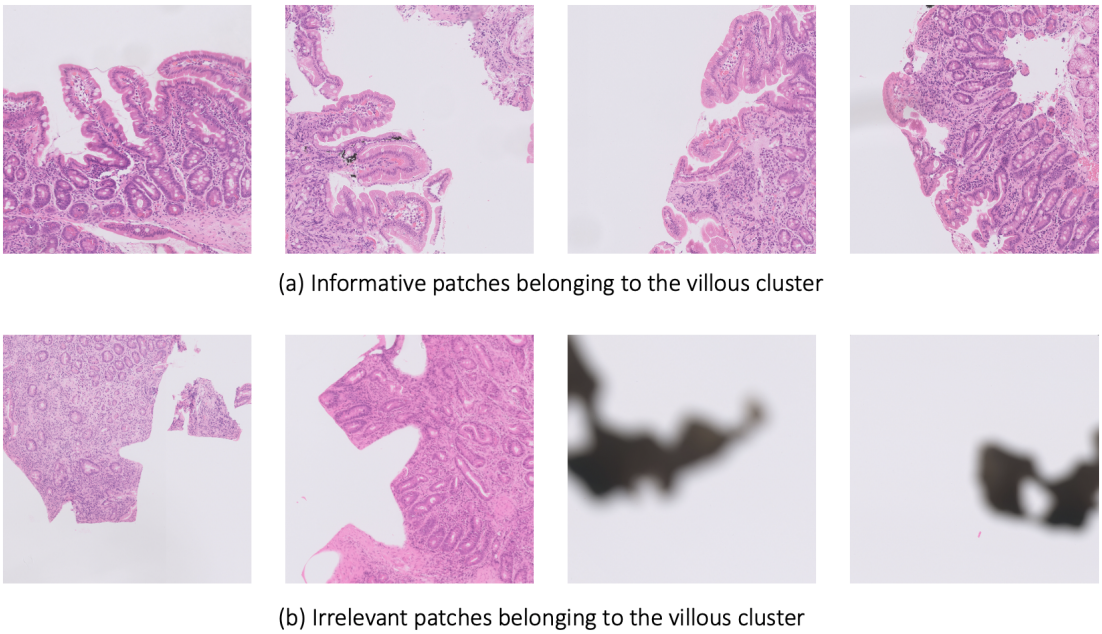


Figure 6.1: Sample patches belonging to the cluster of villi. Most of the prominent patches in the villous region were effectively identified, however, some irrelevant patches containing artifacts or square outlines also belonged to the villous cluster.

patches and build a more informative patch-level training set along with saving the cost of region-based annotations.

Some of the research discussed in this thesis has already considered this aspect. In section 5.3.4, a simple yet efficient preprocessing pipeline was used to pick up informative patches from the coeliac disease dataset in order to train the proposed patch-level classification network more efficiently. The strategy used a convolutional autoencoder to extract the high-level embedded features and then used a k-means algorithm for patch clustering. This was chosen because the distinguishing features for the diagnosis of coeliac disease severity can be found in separate and clearly defined regions. To evaluate the effectiveness of this preprocessing pipeline, two patch-based datasets were constructed. In the first dataset 1000 patches of size 512×512 pixels were randomly cropped from each whole-slide images. In the second dataset the centres of each patches belonging to the cluster of villi were extracted to generate weakly annotated data points. As a result, a large number of 512×512 patches around the annotation points were cut for patch-level training. It can be found that on the dataset with random cropping, the standard Resnet yields a

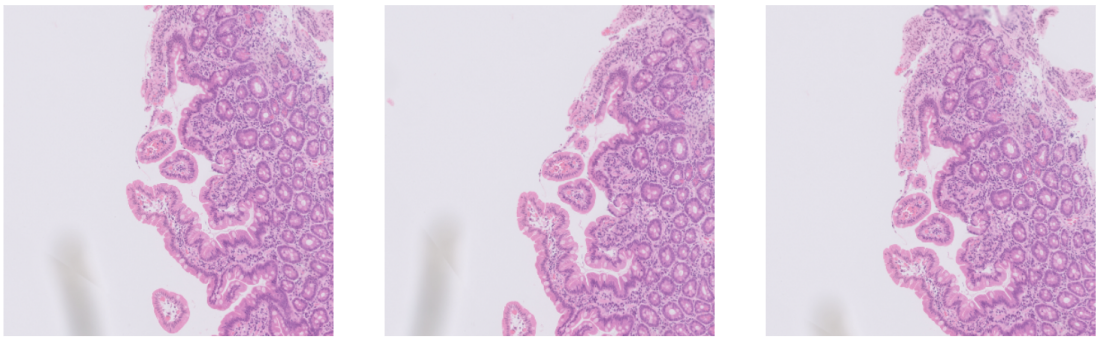


Figure 6.2: Three patches selected from the same whole-slide image. The preprocessing process may lead to a large number of redundant patches in the patch-level dataset.

classification accuracy of 51.2%, whereas when the dataset with villous detection is used, the accuracy improved to 76.6% using the same Resnet backbone. This result indicates that the majority of the salient regions were identified, however, some useless patches also fell into the villous cluster, as shown in Figure 6.1. Moreover, the whole-slide images were split into small segments (128×128 pixels) for two-stage clustering as its scale is similar to the width of the villous extrusion, and in the subsequent patch-level classification, large patches of size 1024×1024 pixels were extracted around the villous segments to include more spatial contextual information. This process leads to a large number of redundant patches in the patch-level dataset (Figure 6.2), which reduces the efficiency of network training.

In the next step, the aim would be to design an automatic patch location system to build representative training instances. There are some directions that can be followed on how to select the most informative regions for whole-slide image classification tasks. For example, active learning-based methods have provided the opportunity to select representative or informative instances from the training set. Instance information means that each instance in the training set brings different information to the model training, i.e., each instance contributes to the model training in a different way. The key idea of active learning is that a deep learning-based model uses a certain algorithm to repeatedly search for the most useful unlabelled instances, which are then labelled by experts and added to the

labelled dataset until the training process converges. As a result, if it can select the representative instances that are most beneficial to the learning process, it can achieve the desired performance with fewer labelled training instances. In [182], Yang et al. present a deep active learning framework that combines fully convolutional networks and active learning to reduce the annotation effort for histological image segmentation tasks. This approach starts with very little training data and repeatedly trains a deep network. At the end of each phase, it extracts useful information (e.g. the most effective areas) from the network to determine the next batch of images to annotate. After acquiring new annotation data, the next phase is started using all available annotated images. Therefore, following this line of research, it was natural to explore whether these active learning-based methods are able to select the most representative spots in the region of interest for the classification tasks of whole-slide images. This could be used to create an automatic annotator that can be trained from a limited number of region-based expert annotations and then learn to automatically annotate the most discriminating regions for incoming new whole-slide images.

6.2.2 Weakly Supervised Whole-slide Image Classification

Automatic classification of whole-slide histopathological images is a fundamental step in the development of computational pathology tools. In clinical practice, pathologists routinely perform histological assessments by inspecting certain regions/components in whole-slide images for better cancer grading and diagnosis. However, it usually requires more effort and expense to obtain region-based annotations from a group of broad-certificated histopathology professionals, thus requiring weakly supervised approaches that allow the classification tasks to be carried out with limited annotations. To achieve this, the network design needs to consider two levels: patch-level feature learning and image-level aggregation.

This thesis presented an attention-based multi-instance classification framework that has the potential to serve as a basis for patch-level classification. However, it remains an open question on how to improve the consistency and discrimination

of the learned patch-level feature representations. In Section 5.2.3, inspired by knowledge distillation strategy [167], I proposed an instance representation loss to enforce the network to learn the consistent features across the most and less representative patches sampled from the same image. On top of that, there is a number of methods in the literature that addresses similar questions. For example, Venkateswara et al. [183] proposed the Maximum Mean Discrepancy (MMD) loss to reduce the inter-domain discrepancy between the source and target feature representations. This loss function is also based on the same assumption that if two images have the same label, then their feature representations should be similar. Xu et al. [184] use a Graph Neural Network module to consider the higher-order connection between images, and further enforce the images with the same label to have similar feature representations. In particular, they first construct a graph using the features extracted from the backbone CNN, with each vertex representing the features from the source images and the target images. There is an edge between the two pairs of vertices only when the distance between the corresponding features is less than a specific threshold, followed by feeding the image representations into a graph neural network to smooth and align the representations in the feature space. These methods can be incorporated into multiple layers of the network to minimise the differences in feat space between images with the same label, however, the instance representation loss can only be applied to the output of the softmax layers. It would be an interesting extension if we could investigate different ways to exploit the feature learning capabilities of deep neural networks in the future.

Furthermore, there is still some room for improvement in the aggregation of patch-level information in whole-slide image analysis. Since the coeliac disease dataset is only recently available, in this thesis I simply combined the patch-level results based on majority voting, where the category with the highest frequency of patch-level predictions was chosen as the whole slide severity diagnosis for coeliac disease. In the discussion of some failure cases, it turned out that the algorithm had difficulty estimating the contribution of different patches to image-level labels and failed to eliminate the confounding effect of some useless patches. In the

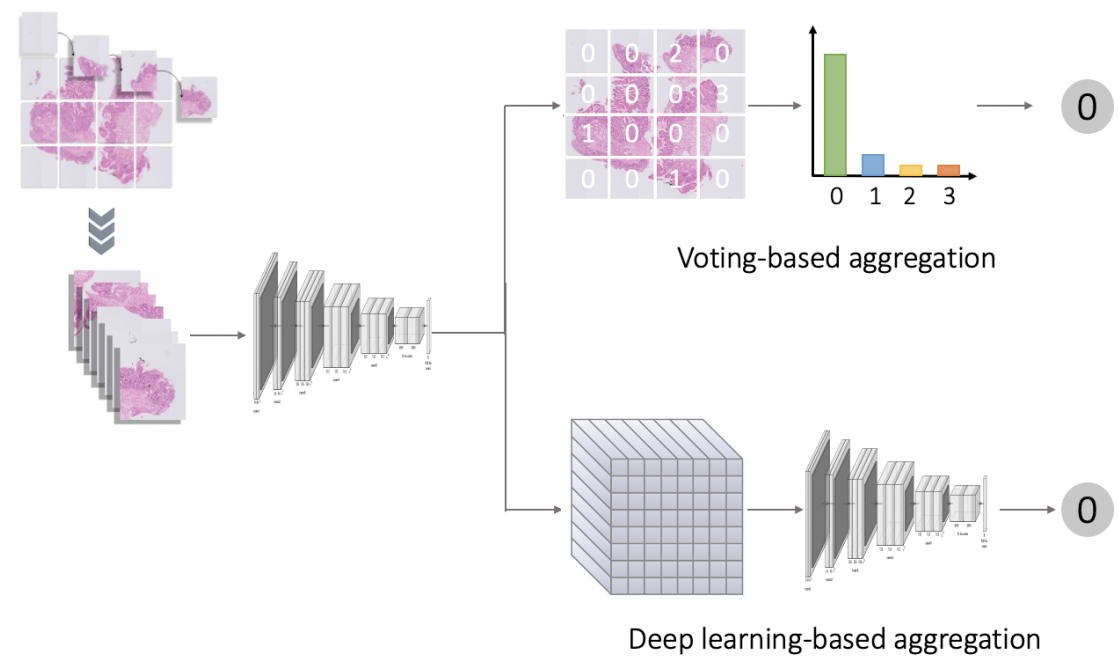


Figure 6.3: A comparison of voting-based and deep learning-based feature aggregation. In contrast to voting-based methods that aggregate on the patch-level classification results, deep learning-based methods extracts the high-dimensional feature vectors and feeds them as input to a second deep network for image-level prediction

last few years, deep neural networks, especially convolutional neural networks (CNNs), have become mainstream tools for a number of applications in biomedical image analysis. Thus, a large body of work has applied deep learning techniques to aggregate these patch-based local representations. In contrast to voting-based methods that aggregate on the patch-level classification results, this class of methods extracts the learned high-dimensional feature vectors from a patch classifier and feeds them as input to a second deep network for image-level prediction (Figure 6.3). Shaban et al. [97] leveraged three different CNN-based architectures, including conventional convolutional blocks, residual blocks and pyramid blocks, to aggregate the local representations for final image-level prediction. The results of these deep architectures with different complexity and capabilities demonstrated the ability of the proposed framework to learn large contextual information and achieve state-of-the-art classification performance. Wang et al. [135] designed a deep network equipped with an attention mechanism to aggregate the patch-based features for image-level prediction. The method noted that each instance contributes

differently to the final image-level prediction based on its inherent discriminability, and therefore they proposed an instance recalibration module that recalibrated the patch-level features based on the importance factors computed by fusing local and global features, which made the whole-slide classification more robust even if some irrelevant patch were previously selected for patch-based training. Future research might investigate how to aggregate the patch-level information for a better image-level prediction.

6.3 Summary

The research presented in this thesis explored traditional and deep learning-based techniques for fine-grained biomedical image analysis. Relative to other publications [185], the classification performance of the currently proposed method is acceptable for clinical assessment and has the potential to provide a reliable diagnosis in clinical practice. I also believe that the contributions of the Clinical Fungal Dataset and the coeliac disease dataset will be constructive to the research community in biomedical image analysis. In addition to investigating the technical details of whole-slide image analysis, future work will aim to further examine the performance of the proposed method in other independent clinical cohorts and thus facilitate the application of the technique to routine clinical examinations. Specifically, two external independent cohort datasets for coeliac disease are currently being constructed and will be used in future clinical studies to improve the diagnosis of coeliac disease. It is anticipated that the proposed approaches will assist pathologists in the future to make accurate and effective diagnoses in their daily workflow.

References

[1] Weidi Xie, J Alison Noble, and Andrew Zisserman. "Microscopy cell counting and detection with fully convolutional regression networks". In: *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 6.3 (2018), pp. 283–292.

[2] Ismael Kone and Lahsen Boulmane. "Hierarchical resnet models for breast cancer histology image classification". In: *International Conference Image Analysis and Recognition*. Springer. 2018, pp. 796–803.

[3] Hoo-Chang Shin et al. "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning". In: *IEEE Transactions on Medical Imaging* 35.5 (2016), pp. 1285–1298.

[4] Sehwan Ki et al. "Fully end-to-end learning based conditional boundary equilibrium GAN with receptive field sizes enlarged for single ultra-high resolution image dehazing". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018, pp. 817–824.

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2015, pp. 234–241.

[6] N Ramesh, J-H Yoo, and IK Sethi. "Thresholding based on histogram approximation". In: *IEE Proceedings-Vision, Image and Signal Processing* 142.5 (1995), pp. 271–279.

[7] TW Ridler, S Calvard, et al. "Picture thresholding using an iterative selection method". In: *IEEE Trans Syst Man Cybern* 8.8 (1978), pp. 630–632.

[8] Jagat Narain Kapur, Prasanna K Sahoo, and Andrew KC Wong. "A new method for gray-level picture thresholding using the entropy of the histogram". In: *Computer Vision, Graphics, and Image Processing* 29.3 (1985), pp. 273–285.

[9] Josef Kittler and John Illingworth. "Minimum error thresholding". In: *Pattern Recognition* 19.1 (1986), pp. 41–47.

[10] H-S Wu, J Berba, and J Gil. "Iterative thresholding for segmentation of cells from noisy images". In: *Journal of Microscopy* 197.3 (2000), pp. 296–304.

[11] Boqiang Liu et al. "Microscopic image analysis and recognition on pathological cells". In: *Canadian Conference on Electrical and Computer Engineering*. IEEE. 2007, pp. 1022–1025.

[12] Archana A Nawandhar, Lakshmi Yamujala, and Navin Kumar. "Image segmentation using thresholding for cell nuclei detection of colon tissue". In: *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE. 2015, pp. 1199–1203.

[13] Rolf Adams and Leanne Bischof. "Seeded region growing". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16.6 (1994), pp. 641–647.

[14] Jianping Fan et al. "Automatic image segmentation by integrating color-edge extraction and seeded region growing". In: *IEEE Transactions on Image Processing* 10.10 (2001), pp. 1454–1466.

[15] Frank Y Shih and Shouxian Cheng. "Automatic seeded region growing for color image segmentation". In: *Image and Vision Computing* 23.10 (2005), pp. 877–886.

[16] Andrew Mehnert and Paul Jackway. "An improved seeded region growing algorithm". In: *Pattern Recognition Letters* 18.10 (1997), pp. 1065–1071.

[17] SA Hojjatoleslami and Josef Kittler. "Region growing: a new approach". In: *IEEE Transactions on Image Processing* 7.7 (1998), pp. 1079–1084.

[18] S-Y Wan and William E Higgins. "Symmetric region growing". In: *IEEE Transactions on Image Processing* 12.9 (2003), pp. 1007–1015.

[19] Dwi Anoraganingrum, Sabine Kroner, and Bjorn Gottfried. "Cell segmentation with adaptive region growing". In: *ICIAP Venedig, Italy* (1999), pp. 27–29.

[20] Sophie Liu Xiao Fan, Zhihong Man, and Rashi Samur. "Edge based region growing: a new image segmentation method". In: *Proceedings of the 2004 ACM SIGGRAPH International Conference on Virtual Reality Continuum and Its Applications in Industry*. 2004, pp. 302–305.

[21] Jamshid Dehmeshki, Xujiong Ye, and John Costello. "Shape based region growing using derivatives of 3D medical images: application to semiautomated detection of pulmonary nodules". In: *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)*. Vol. 1. IEEE. 2003, pp. I–1085.

[22] Xiaodong Yang, Houqiang Li, and Xiaobo Zhou. "Nuclei segmentation using marker-controlled watershed, tracking using mean-shift, and Kalman filter in time-lapse microscopy". In: *IEEE Transactions on Circuits and Systems* 53.11 (2006), pp. 2405–2414.

[23] PS Umesh Adiga and BB Chaudhuri. "An efficient method based on watershed and rule-based merging for segmentation of 3-D histo-pathological images". In: *Pattern Recognition* 34.7 (2001), pp. 1449–1458.

[24] Carlos Ortiz De Solorzano et al. "Segmentation of confocal microscope images of cell nuclei in thick tissue sections". In: *Journal of Microscopy* 193.3 (1999), pp. 212–226.

[25] Sandor Szenasi. "Medical image segmentation with split-and-merge method". In: *5th IEEE International Symposium on Logistics and Industrial Informatics (LINDI 2013)*. 2013, pp. 137–140.

[26] Yunyong Surut and Pornchai Phukpattaranont. "Overlapping cell image segmentation using surface splitting and surface merging algorithms". In: *Second APSIPA Annual Summit and Conference*. 2010, pp. 662–666.

[27] Radhakrishna Achanta et al. "SLIC superpixels compared to state-of-the-art superpixel methods". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.11 (2012), pp. 2274–2282.

[28] Chi-Yu Hsu and Jian-Jiun Ding. "Efficient image segmentation algorithm using SLIC superpixels and boundary-focused region merging". In: *International Conference on Information, Communications and Signal Processing*. IEEE. 2013, pp. 1–5.

[29] Jianbo Shi and Jitendra Malik. "Normalized cuts and image segmentation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.8 (2000), pp. 888–905.

[30] Timothee Cour, Florence Benezit, and Jianbo Shi. "Spectral segmentation with multiscale graph decomposition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Vol. 2. IEEE. 2005, pp. 1124–1131.

[31] Yong He et al. "iCut: an integrative cut algorithm enables accurate segmentation of touching cells". In: *Scientific Reports* 5 (2015).

[32] Yousef Al-Kofahi et al. "Improved automatic detection and segmentation of cell nuclei in histopathology images". In: *IEEE Transactions on Biomedical Engineering* 57.4 (2010), pp. 841–852.

[33] John Canny. "A computational approach to edge detection". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6 (1986), pp. 679–698.

[34] Catherine Garbay. "Image structure representation and processing: a discussion of some segmentation methods in cytology". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2 (1986), pp. 140–146.

[35] Dana H Ballard. "Generalizing the Hough transform to detect arbitrary shapes". In: *Pattern Recognition* 13.2 (1981), pp. 111–122.

[36] Eric Cosatto et al. "Grading nuclear pleomorphism on histological micrographs". In: *International Conference on Pattern Recognition*. IEEE. 2008, pp. 1–4.

[37] Carlos Ortiz De Solorzano et al. "Segmentation of confocal microscope images of cell nuclei in thick tissue sections". In: *Journal of Microscopy* 193.3 (1999), pp. 212–226.

[38] C Ortiz de Solorzano et al. "Segmentation of nuclei and cells using membrane related protein markers". In: *Journal of Microscopy* 201.3 (2001), pp. 404–415.

[39] Chi Liu et al. "Detecting and segmenting cell nuclei in two-dimensional microscopy images". In: *Journal of Pathology Informatics* 7 (2016).

[40] Alan L Yuille, Peter W Hallinan, and David S Cohen. "Feature extraction from faces using deformable templates". In: *International Journal of Computer Vision* 8.2 (1992), pp. 99–111.

[41] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. "Snakes: Active contour models". In: *International Journal of Computer Vision* 1.4 (1988), pp. 321–331.

[42] Ravi Malladi, James A Sethian, and Baba C Vemuri. "Shape modeling with front propagation: A level set approach". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 17.2 (1995), pp. 158–175.

[43] Vicent Caselles, Ron Kimmel, and Guillermo Sapiro. "Geodesic active contours". In: *International Journal of Computer Vision* 22.1 (1997), pp. 61–79.

[44] Hongmei Zhang et al. "An efficient multiscale approach to level set evolution". In: *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No. 03CH37439)*. Vol. 1. IEEE. 2003, pp. 694–697.

[45] Bo Zhang, Christophe Zimmer, and J-C Olivo-Marin. "Tracking fluorescent cells with coupled geometric active contours". In: *2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro (IEEE Cat No. 04EX821)*. IEEE. 2004, pp. 476–479.

[46] Alexandre Dufour et al. "Segmenting and tracking fluorescent cells in dynamic 3-D microscopy with coupled active surfaces". In: *IEEE Transactions on Image Processing* 14.9 (2005), pp. 1396–1410.

[47] Thomas Nketia et al. "Analysis of live cell images: Methods, tools and opportunities". In: *Methods* (2017).

[48] Kaustav Nandy et al. "Automatic segmentation and supervised learning-based selection of nuclei in cancer tissue images". In: *Cytometry Part A* 81.9 (2012), pp. 743–754.

[49] Christophe Zimmer et al. "Segmentation and tracking of migrating cells in videomicroscopy with parametric active contours: A tool for cell-based drug testing". In: *IEEE Transactions on Medical Imaging* 21.10 (2002), pp. 1212–1221.

[50] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[51] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[52] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 3431–3440.

[53] Kaiming He et al. "Mask r-cnn". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2961–2969.

[54] Tsung-Yi Lin et al. "Feature pyramid networks for object detection". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 2117–2125.

[55] Tianyi Zhao and Zhaozheng Yin. "Pyramid-based fully convolutional networks for cell segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 677–685.

[56] Chengjia Wang et al. "A two-stage 3D Unet framework for multi-class segmentation on full resolution image". In: *arXiv preprint arXiv:1804.04341* (2018).

[57] Anton Bohm et al. "ISOO DL: Instance segmentation of overlapping biological objects using deep learning". In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE. 2018, pp. 1225–1229.

[58] Shan E Ahmed Raza et al. "Micro-Net: A unified model for segmentation of various objects in microscopy images". In: *Medical Image Analysis* 52 (2019), pp. 160–173.

[59] Zongwei Zhou et al. “Unet++: Redesigning skip connections to exploit multiscale features in image segmentation”. In: *IEEE Transactions on Medical Imaging* 39.6 (2019), pp. 1856–1867.

[60] Neeraj Kumar et al. “A dataset and a technique for generalized nuclear segmentation for computational pathology”. In: *IEEE Transactions on Medical Imaging* 36.7 (2017), pp. 1550–1560.

[61] Hao Chen et al. “DCAN: deep contour-aware networks for accurate gland segmentation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 2487–2496.

[62] Hirohisa Oda et al. “BESNet: boundary-enhanced segmentation of cells in histopathological images”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 228–236.

[63] Yanning Zhou et al. “Cia-net: Robust nuclei instance segmentation with contour-aware information aggregation”. In: *International Conference on Information Processing in Medical Imaging*. Springer. 2019, pp. 682–693.

[64] Mina Khoshdeli and Bahram Parvin. “Deep learning models delineates multiple nuclear phenotypes in h&e stained histology sections”. In: *arXiv preprint arXiv:1802.04427* (2018).

[65] Wenjia Bai et al. “Semi-supervised learning for network-based cardiac MR image segmentation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 253–260.

[66] Lin Gu et al. “Semi-supervised learning for biomedical image segmentation via forest oriented super pixels (voxels)”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 702–710.

[67] Nima Tajbakhsh et al. “Surrogate supervision for medical image analysis: Effective deep learning from limited quantities of labeled data”. In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE. 2019, pp. 1251–1255.

[68] Wenjia Bai et al. “Self-supervised learning for cardiac mr image segmentation by anatomical position prediction”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 541–549.

[69] Geoffrey E Hinton et al. “Improving neural networks by preventing co-adaptation of feature detectors”. In: *arXiv preprint arXiv:1207.0580* (2012).

[70] Christian Szegedy et al. “Going deeper with convolutions”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 1–9.

[71] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 770–778.

[72] Hongliu Cao et al. “Improve the performance of transfer learning without fine-tuning using dissimilarity-based multi-view learning for breast cancer histology images”. In: *International Conference Image Analysis and Recognition*. Springer. 2018, pp. 779–787.

[73] Yeeleng S Vang, Zhen Chen, and Xiaohui Xie. "Deep learning framework for multi-class breast cancer histology image classification". In: *International Conference Image Analysis and Recognition*. Springer. 2018, pp. 914–922.

[74] Aditya Golatkar, Deepak Anand, and Amit Sethi. "Classification of breast cancer histology using deep learning". In: *International Conference Image Analysis and Recognition*. Springer. 2018, pp. 837–844.

[75] Rui Yan et al. "Breast cancer histopathological image classification using a hybrid deep neural network". In: *Methods* (2019).

[76] Jiawen Yang et al. "Classification for dermoscopy images using convolutional neural networks based on region average pooling". In: *IEEE Access* 6 (2018), pp. 65130–65138.

[77] Nazneen N Sultana, Bappaditya Mandal, and Niladri B Puhan. "Deep residual network with regularised fisher framework for detection of melanoma". In: *IET Computer Vision* 12.8 (2018), pp. 1096–1104.

[78] Jianpeng Zhang et al. "Attention residual learning for skin lesion classification". In: *IEEE Transactions on Medical Imaging* (2019).

[79] Wei Shen et al. "Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification". In: *Pattern Recognition* 61 (2017), pp. 663–673.

[80] Yutong Xie et al. "Fusing texture, shape and deep model-learned information at decision level for automated classification of lung nodules on chest CT". In: *Information Fusion* 42 (2018), pp. 102–110.

[81] Hui Chen et al. "Classification of pulmonary nodules using neural network ensemble". In: *International Symposium on Neural Networks*. Springer. 2011, pp. 460–466.

[82] Alexander Rakhlin et al. "Deep convolutional neural networks for breast cancer histology image analysis". In: *International Conference Image Analysis and Recognition*. Springer. 2018, pp. 737–744.

[83] Carlos A Ferreira et al. "Classification of breast cancer histology images through transfer learning using a pre-trained inception resnet v2". In: *International Conference Image Analysis and Recognition*. Springer. 2018, pp. 763–770.

[84] Nadia Brancati, Maria Frucci, and Daniel Riccio. "Multi-classification of breast cancer histology images by using a fine-tuning strategy". In: *International Conference Image Analysis and Recognition*. Springer. 2018, pp. 771–778.

[85] Andre Esteva et al. "Dermatologist-level classification of skin cancer with deep neural networks". In: *Nature* 542.7639 (2017), p. 115.

[86] Kazuhisa Matsunaga et al. "Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble". In: *arXiv preprint arXiv:1703.03108* (2017).

[87] Balazs Harangi. "Skin lesion classification with ensembles of deep convolutional neural networks". In: *Journal of Biomedical Informatics* 86 (2018), pp. 25–32.

[88] Lei Bi et al. "Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks". In: *arXiv preprint arXiv:1703.04197* (2017).

[89] Terrance DeVries and Dhanesh Ramachandram. "Skin lesion classification using deep multi-scale convolutional neural networks". In: *arXiv preprint arXiv:1703.01402* (2017).

[90] Noel CF Codella et al. "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)". In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE. 2018, pp. 168–172.

[91] Zongyuan Ge et al. "Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 250–258.

[92] Afonso Menegola et al. "RECOD titans at ISIC challenge 2017". In: *arXiv preprint arXiv:1703.04819* (2017).

[93] Mingchen Gao et al. "Holistic classification of CT attenuation patterns for interstitial lung diseases via deep convolutional neural networks". In: *Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization* 6.1 (2018), pp. 1–6.

[94] Yiming Lei et al. "Shape and margin-aware lung nodule classification in low-dose CT images via soft activation mapping". In: *Medical Image Analysis* 60 (2020), p. 101628.

[95] Samuel G Armato III et al. "The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans". In: *Medical Physics* 38.2 (2011), pp. 915–931.

[96] Kamyar Nazeri, Azad Aminpour, and Mehran Ebrahimi. "Two-stage convolutional neural network for breast cancer histology image classification". In: *International Conference Image Analysis and Recognition*. Springer. 2018, pp. 717–726.

[97] Muhammad Shaban et al. "Context-aware convolutional neural network for grading of colorectal cancer histology images". In: *IEEE Transactions on Medical Imaging* (2020).

[98] Nils Gessert et al. "Skin lesion classification using cnns with patch-based attention and diagnosis-guided loss weighting". In: *IEEE Transactions on Biomedical Engineering* 67.2 (2019), pp. 495–503.

[99] Ivan Gonzalez Diaz. "Incorporating the knowledge of dermatologists to convolutional neural networks for the diagnosis of skin lesions". In: *arXiv preprint arXiv:1703.01976* (2017).

[100] Thomas G Dietterich, Richard H Lathrop, and Tomas Lozano-Perez. "Solving the multiple instance problem with axis-parallel rectangles". In: *Artificial Intelligence* 89.1-2 (1997), pp. 31–71.

[101] Wentao Zhu et al. "Deep multi-instance networks with sparse label assignment for whole mammogram classification". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 603–611.

[102] Oren Z Kraus, Jimmy Lei Ba, and Brendan J Frey. "Classifying and segmenting microscopy images with deep multiple instance learning". In: *Bioinformatics* 32.12 (2016), pp. i52–i59.

[103] Le Hou et al. "Patch-based convolutional neural network for whole slide tissue image classification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 2424–2433.

[104] Zhennan Yan et al. "Multi-instance deep learning: Discover discriminative local anatomies for bodypart recognition". In: *IEEE Transactions on Medical Imaging* 35.5 (2016), pp. 1332–1343.

[105] Renato Hermoza et al. "Region Proposals for Saliency Map Refinement for Weakly-Supervised Disease Localisation and Classification". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 539–549.

[106] Bolei Zhou et al. "Learning deep features for discriminative localization". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 2921–2929.

[107] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps". In: *arXiv preprint arXiv:1312.6034* (2013).

[108] Guilin Liu et al. "Image inpainting for irregular holes using partial convolutions". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 85–100.

[109] Ashish Vaswani et al. "Attention is all you need". In: *Advances in Neural Information Processing Systems*. 2017, pp. 5998–6008.

[110] Jun Fu et al. "Dual attention network for scene segmentation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 3146–3154.

[111] Lianli Gao et al. "Video captioning with attention-based LSTM and semantic consistency". In: *IEEE Transactions on Multimedia* 19.9 (2017), pp. 2045–2055.

[112] Pau Rodriguez et al. "Attend and rectify: a gated attention mechanism for fine-grained recovery". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 349–364.

[113] Fei Wang et al. "Residual attention network for image classification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 3156–3164.

[114] Jie Hu, Li Shen, and Gang Sun. "Squeeze-and-excitation networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7132–7141.

[115] Sanghyun Woo et al. "Cbam: Convolutional block attention module". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 3–19.

[116] Jie Hu et al. "Gather-excite: Exploiting feature context in convolutional neural networks". In: *NeurIPS*. 2018, pp. 9401–9411.

[117] Xiaolong Wang et al. "Non-local neural networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7794–7803.

[118] Zilong Huang et al. "Ccnet: Criss-cross attention for semantic segmentation". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pp. 603–612.

[119] Heliang Zheng et al. "Looking for the Devil in the Details: Learning Trilinear Attention Sampling Network for Fine-grained Image Recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 5012–5021.

[120] Yue Cao et al. "GCNet: Non-local Networks Meet Squeeze-Excitation Networks and Beyond". In: *arXiv preprint arXiv:1904.11492* (2019).

[121] Qingji Guan et al. "Thorax disease classification with attention guided convolutional neural network". In: *Pattern Recognition Letters* 131 (2020), pp. 38–45.

[122] Yixin Wang et al. "FGB: Feature Guidance Branch for Organ Detection in Medical Images". In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2020, pp. 349–353.

[123] Shuai Chen et al. "Multi-task attention-based semi-supervised learning for medical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 457–465.

[124] Jo Schlemper et al. "Attention gated networks: Learning to leverage salient regions in medical images". In: *Medical Image Analysis* 53 (2019), pp. 197–207.

[125] Heechan Yang et al. "Guided soft attention network for classification of breast cancer histopathology images". In: *IEEE Transactions on Medical Imaging* 39.5 (2019), pp. 1306–1315.

[126] Jaemin Son et al. "Classification of findings with localized lesions in fundoscopic images using a regionally guided CNN". In: *Computational Pathology and Ophthalmic Medical Image Analysis*. Springer, 2018, pp. 176–184.

[127] Jingru Yi et al. "Attentive neural cell instance segmentation". In: *Medical Image Analysis* 55 (2019), pp. 228–240.

[128] Ashish Sinha and Jose Dolz. "Multi-scale guided attention for medical image segmentation". In: *arXiv preprint arXiv:1906.02849* (2019).

[129] Yan Wang et al. "Abdominal multi-organ segmentation with organ-attention networks and statistical fusion". In: *Medical Image Analysis* 55 (2019), pp. 88–102.

[130] Yi Wang et al. "Deep attentive features for prostate segmentation in 3d transrectal ultrasound". In: *IEEE Transactions on Medical Imaging* 38.12 (2019), pp. 2768–2778.

[131] Jingyi Wang et al. "AEC-Net: Attention and Edge Constraint Network for Medical Image Segmentation". In: *2020 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2020, pp. 1616–1619.

[132] Caizi Li et al. "Attention based hierarchical aggregation network for 3D left atrial segmentation". In: *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer. 2018, pp. 255–264.

[133] Maximilian Ilse, Jakub M Tomczak, and Max Welling. "Attention-based deep multiple instance learning". In: *arXiv preprint arXiv:1802.04712* (2018).

[134] Ming Y Lu et al. "Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding". In: *arXiv preprint arXiv:1910.10825* (2019).

[135] Shujun Wang et al. "RMDL: Recalibrated multi-instance deep learning for whole slide gastric image classification". In: *Medical Image Analysis* 58 (2019), p. 101549.

[136] Dawid Rymarczyk, Jacek Tabor, and Bartosz Zielinski. "Kernel Self-Attention in Deep Multiple Instance Learning". In: *arXiv preprint arXiv:2005.12991* (2020).

[137] Jiongxin Liu et al. "Dog breed classification using part localization". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer. 2012, pp. 172–185.

[138] Thomas Berg and Peter N Belhumeur. "Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013, pp. 955–962.

[139] Lingxi Xie et al. "Hierarchical part matching for fine-grained visual categorization". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2013, pp. 1641–1648.

[140] Steve Branson et al. "Bird species categorization using pose normalized deep convolutional nets". In: *arXiv preprint arXiv:1406.2952* (2014).

[141] Han Zhang et al. "Spda-cnn: Unifying semantic part detection and abstraction for fine-grained recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 1143–1152.

[142] Yuning Chai, Victor Lempitsky, and Andrew Zisserman. "Symbiotic segmentation and part localization for fine-grained categorization". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2013, pp. 321–328.

[143] Ning Zhang et al. "Panda: Pose aligned networks for deep attribute modeling". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 1637–1644.

[144] Ning Zhang et al. "Deformable part descriptors for fine-grained recognition and attribute prediction". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2013, pp. 729–736.

[145] Ning Zhang et al. "Part-based R-CNNs for fine-grained category detection". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer. 2014, pp. 834–849.

[146] Jonathan Krause et al. "Fine-grained recognition without part annotations". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 5546–5555.

[147] Tianjun Xiao et al. "The application of two-level attention models in deep convolutional neural network for fine-grained image classification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 842–850.

[148] Jianlong Fu, Heliang Zheng, and Tao Mei. "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 4438–4446.

[149] Ming Sun et al. "Multi-attention multi-class constraint for fine-grained image recognition". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 805–821.

[150] Heliang Zheng et al. "Learning multi-attention convolutional neural network for fine-grained image recognition". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 5209–5217.

[151] Bo Zhao et al. "Diversified visual attention networks for fine-grained object classification". In: *IEEE Transactions on Multimedia* 19.6 (2017), pp. 1245–1256.

[152] Yu Zhang et al. "Weakly supervised fine-grained categorization with part-based image representation". In: *IEEE Transactions on Image Processing* 25.4 (2016), pp. 1713–1725.

[153] Ze Yang et al. "Learning to navigate for fine-grained classification". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 420–435.

[154] Weifeng Ge, Xiangru Lin, and Yizhou Yu. "Weakly supervised complementary parts models for fine-grained image classification from the bottom up". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 3034–3043.

[155] Charles Sutton and Andrew McCallum. "An introduction to conditional random fields for relational learning". In: *Introduction to Statistical Relational Learning* 2 (2006), pp. 93–128.

[156] Sepp Hochreiter and Jurgen Schmidhuber. "Long short-term memory". In: *Neural Computation* 9.8 (1997), pp. 1735–1780.

[157] Jens Rittscher, Peter H Tu, and Nils Krahnstoever. "Simultaneous estimation of segmentation and shape". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Vol. 2. IEEE. 2005, pp. 486–493.

[158] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. "Learning deconvolution network for semantic segmentation". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 1520–1528.

[159] John Canny. "A computational approach to edge detection". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6 (1986), pp. 679–698.

[160] Jianpeng Zhang et al. "Medical image classification using synergic deep learning". In: *Medical Image Analysis* 54 (2019), pp. 10–19.

[161] Catherine Wah et al. *The caltech-ucsd birds-200-2011 dataset*. Tech. rep. CNS-TR-2011-001. California Institute of Technology, 2011.

[162] Jonathan Krause et al. "3d object representations for fine-grained categorization". In: *Proceedings of the IEEE International Conference on Computer Vision workshops*. 2013, pp. 554–561.

[163] Luc Devroye. "Sample-based non-uniform random variate generation". In: *Proceedings of the 18th Conference on Winter simulation*. 1986, pp. 260–265.

[164] Noor Atiqah Zulkifli and Latiffah Zakaria. "Morphological and molecular diversity of Aspergillus from corn grain used as livestock feed". In: *HAYATI Journal of Biosciences* 24.1 (2017), pp. 26–34.

[165] K Diba et al. "Identification of Aspergillus species using morphological characteristics". In: *Pakistan Journal of Medical Sciences* 23.6 (2007), p. 867.

[166] Liang-Chieh Chen et al. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.4 (2017), pp. 834–848.

[167] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network". In: *arXiv preprint arXiv:1503.02531* (2015).

[168] Gershon Buchsbaum. "A spatial processor model for object colour perception". In: *Journal of the Franklin Institute* 310.1 (1980), pp. 1–26.

[169] Alessio Fasano and Carlo Catassi. "Current approaches to diagnosis and treatment of celiac disease: an evolving spectrum". In: *Gastroenterology* 120.3 (2001), pp. 636–651.

[170] Carlo Catassi and Alessio Fasano. "Celiac disease diagnosis: simple rules are better than complicated algorithms". In: *The American Journal of Medicine* 123.8 (2010), pp. 691–693.

[171] Ilaria Parzanese et al. "Celiac disease: From pathophysiology to treatment". In: *World Journal of Gastrointestinal Pathophysiology* 8.2 (2017), p. 27.

[172] Gino Roberto Corazza et al. "Comparison of the interobserver reproducibility with different histologic criteria used in celiac disease". In: *Clinical Gastroenterology and Hepatology* 5.7 (2007), pp. 838–843.

[173] Michael N Marsh. "Gluten, major histocompatibility complex, and the small intestine: a molecular and immunobiologic approach to the spectrum of gluten sensitivity ('celiac sprue')". In: *Gastroenterology* 102.1 (1992), pp. 330–354.

[174] Georg Oberhuber, Gerhard Granditsch, and Harald Vogelsang. "The histopathology of coeliac disease: time for a standardized report scheme for pathologists". In: *European Journal of Gastroenterology and Hepatology* 11.10 (1999), p. 1185.

[175] Mengran Fan et al. "Fine-Grained Multi-Instance Classification in Microscopy Through Deep Attention". In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2020, pp. 169–173.

[176] Matthias Kohl et al. "Assessment of breast cancer histology using densely connected convolutional networks". In: *International Conference Image Analysis and Recognition*. Springer. 2018, pp. 903–913.

[177] Wajahat Nawaz et al. "Classification of breast cancer histology images using alexnet". In: *International Conference Image Analysis and Recognition*. Springer. 2018, pp. 869–876.

[178] Xulei Yang et al. "A novel multi-task deep learning model for skin lesion segmentation and classification". In: *arXiv preprint arXiv:1703.01025* (2017).

[179] Rafael Teixeira Sousa and Larissa Vasconcellos de Moraes. "Araguaia medical vision lab at ISIC 2017 skin lesion classification challenge". In: *arXiv preprint arXiv:1703.00856* (2017).

[180] Xi Jia and Linlin Shen. "Skin lesion classification using class activation map". In: *arXiv preprint arXiv:1703.01053* (2017).

[181] Hensin Tsao et al. "Early detection of melanoma: reviewing the ABCDEs". In: *Journal of the American Academy of Dermatology* 72.4 (2015), pp. 717–723.

[182] Lin Yang et al. "Suggestive annotation: A deep active learning framework for biomedical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 399–407.

[183] Hemanth Venkateswara et al. "Deep hashing network for unsupervised domain adaptation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 5018–5027.

[184] Dou Xu et al. "Graph Neural Networks for UnsupervisedDomain Adaptation of Histopathological ImageAnalytics". In: *arXiv preprint arXiv:2008.09304* (2020).

[185] Korsuk Sirinukunwattana et al. "Image-based consensus molecular subtype classification (imCMS) of colorectal cancer using deep learning". In: *arXiv. org* 645143 (2019).