POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Deep learning model generalization for microscopy image analysis

DUC HOA TRAN
Département de génie informatique et génie logiciel

Thèse présentée en vue de l'obtention du diplôme de *Philosophiæ Doctor*
Génie informatique

Juin 2022

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Cette thèse intitulée :

Deep learning model generalization for microscopy image analysis

présentée par **Duc Hoa TRAN**
en vue de l'obtention du diplôme de *Philosophiæ Doctor*
a été dûment acceptée par le jury d'examen constitué de :

Guillaume-Alexandre BILODEAU, président
Farida CHERIET, membre et directrice de recherche
Michel MEUNIER, membre et codirecteur de recherche
Julien COHEN-ADAD, membre
Adam KRZYZAK, membre externe

DEDICATION

This thesis is dedicated to my father Van Que Tran, my mother Thi Nhan Tran, my wife Thi Thu Thao Nguyen and my daughter Thao Linh Tran for their love, patience and support to me every moment.

ACKNOWLEDGEMENTS

I'm grateful to my supervisor, Prof. Farida Cheriet for giving me the opportunity to start the Ph.D. program and the freedom to pursue research topics. Above all, I'm grateful to her for providing me with critical guidance and constant support for this research project. I have learned a lot and will always keep her useful advice in mind for my future research.

I'm thankful to my co-supervisor, Prof. Michel Meunier, for transferring his expertise on novel microscopy imaging techniques using plasmonic nanoparticles and his passion for developing microscopy image analysis applications in clinical settings. This research can not be completed without his great support during the last four years.

A huge thanks to Prof. Mohamed Cheriet, who gave out a unique course on Pattern Recognition and Inspection, and brought me the opportunity to complete one of the published articles after the course.

Thank you to Prof. Adam G. Hendricks for valuable lessons on microscopy image acquiring systems and microscopy image analysis. This helped me to have a strong foundation and quickly get familiar with a diverse range of existing microscopes.

Thank you to Prof. Aaron Courville for providing an excellent course on deep representation learning, both in theory and practice. Without it, it would be challenging to catch up with the rapid development of deep learning algorithms.

I would like to thank Cécile Darviot for always being available for support on technical concepts of microscopy imaging techniques and also other non-technical issues when working on the nano-pathological image analysis project. Thank you to Dr. Roula Albadine for the insightful discussion on different aspects of pathological image analysis in clinical settings.

I also want to say thank you to all the members of the LIV4D lab at École Polytechnique of Montreal. Although we have had challenging times due to the pandemic, I always felt grateful for your critical comments and knowledge sharing. A special thank you to Philippe Debanné for his responsible support and availability.

This work was funded by the Quebec's consortium for research and innovation in medical technologies (MEDTEQ), the TransMedTech Institute and the Oncopole. I'm grateful to them when receiving their funding support during the entire research project.

RÉSUMÉ

L’analyse de structures cellulaires dans les images de microscopie est l’une des tâches les plus importantes dans diverses études biologiques et diagnostic de maladies. En particulier, la segmentation et la classification sont deux tâches essentielles qui sont effectuées régulièrement dans la pratique. Cependant, l’hétérogénéité, le débit et la complexité croissants des données générées par les microscopes modernes ont introduit divers défis pour les algorithmes informatiques automatiques. Cela est d’autant plus vrai pour les approches basées sur l’apprentissage profond. Premièrement, la création manuelle d’une vérité terrain suffisante pour entraîner les données dans la classification automatique ou en particulier la segmentation est un défi car elle nécessite une intervention importante de la part des experts du domaine. Ensuite, étant donné le nombre limité d’échantillons de données étiquetées pour l’entrainement, les modèles d’apprentissage profond ont souvent une faible capacité de généralisation. De plus, la complexité élevée du modèle en termes de temps de calcul requis et de processus d’entrainement a un impact sur l’adoption de solutions d’apprentissage profond. Les biologistes ou les pathologistes ont du mal à choisir un modèle d’apprentissage profond existant parmi les autres qui serait plus approprié pour leur application qui nécessite des paramètres spécifiques de préparation d’échantillons et d’acquisition d’images.

L’objectif principal de cette thèse est de développer de nouvelles architectures et techniques d’apprentissage profond pour améliorer la généralisation et l’applicabilité de ces algorithmes pour la segmentation et la classification des images de microscopie dans diverses conditions expérimentales, telles que les types de microscopes, les conditions d’acquisition d’images, les processus de préparation d’échantillons et les catégories de cellules. La première étude examine s’il est possible d’éliminer la dépendance du modèle d’apprentissage profond au processus d’étiquetage manuel pour la segmentation des noyaux dans les images d’histopathologie. Nous avons développé un modèle de segmentation DL non supervisé qui est adaptable au domaine. Notre approche s’est concentrée sur l’exploitation des connaissances préalables des paramètres physiques dans le processus d’acquisition d’images et sur la combinaison d’une architecture d’apprentissage profond avec des algorithmes classiques de traitement d’images. Nous concevons également un module de synthèse de données pour générer des ensembles de données augmentés contenant des exemples similaires à chacune des images cibles tout en identifiant leurs masques correspondants. Les résultats qui ont été obtenus sur trois bases de données publiques d’images histopathologiques démontrent que notre méthode peut surpasser les autres approches de segmentation non supervisée et sa performance est comparable à celle des modèles DL supervisés. Dans les études ultérieures, nous nous concentrons sur la question de savoir s'il est possible de développer des classifieurs basés sur l'apprentissage profond qui ont une bonne généralisation à travers différentes applications étant donné un nombre très limité de données d'entrainement. Nous avons d'abord développé un modèle CNN léger pour classer simultanément les images de microscopie dans plusieurs domaines. Les données cibles comprennent des images qui reperésentent différents niveaux de structures cellulaires, des tissus aux cellules et aux organites cellulaires. De plus, nous élaborons une procédure d'optimisation efficace permettant au réseau proposé de surpasser les méthodes existantes sans nécessiter l'adaptation de paramètres qui sont spécifiques au domaine. Fait intéressant, notre modèle proposé est robuste par rapport à la disponibilité d'un nombre limité de données d'entrainement. Avec une faible complexité et une grande flexibilité en terme d'applicabilité à différents domaines, l'approche devient plus attrayante pour un déploiement dans un contexte de ressources matérielles limitées en laboratoire.

Sur la base du modèle générique précédent, nous avons ensuite exploré la possibilité d'améliorer encore plus la précision d'une application spécifique, en particulier la reconnaissance des organites subcellulaires. Nous avons initialement construit une nouvelle architecture d'apprentissage profond qui combine l'extraction de caractéristiques basée sur DL avec une analyse multi-résolution. Les expériences montrent que le modèle développé peut améliorer considérablement les performances de classification par rapport aux modèles d'apprentissage profond de l'état de l'art sur les mêmes ensembles de données d'images fluorescentes microscopiques. Inspirés par ce modèle, nous avons proposé le développement d'un autre réseau DL compact. Contrairement au modèle précédent dans lequel les couches d'extraction de caractéristiques étaient pré-entrainées à l'aide de l'ensemble populaire de données ImageNet, nous entraînons le nouveau modèle à partir de zéro, sans utiliser d'images externes. Nous avons également formulé un nouveau terme de régularisation à intégrer dans la fonction d'optimisation. Suite à des expériences, les performances du modèle proposé ont dépassé celles des méthodes de l'état de l'art dans des conditions expérimentales similaires et, plus important encore, il peut bien généraliser même avec des données étiquetées très limitées. En particulier, le modèle nécessite environ quatre fois moins de données étiquetées que d'autres approches pour atteindre un niveau de précision similaire. Nous sommes convaincus que les méthodes développées durant ce projet contribueront à réduire considérablement le fardeau de la génération d'annotations de vérité terrain pour les données d'entrainement et à augmenter l'efficacité des pathologistes et des scientifiques biomédicaux dans l'analyse d'images microscopiques. Des travaux futurs devraient se concentrer sur la conception d'une architecture de bout en bout qui peut à la fois effectuer la segmentation et la classification des objets d'intérêt. Une méthode d'apprentissage continu devrait également être investiguée pour aider l'agent d'apprentissage à être en mesure de s'adapter aux nouvelles tâches en fonction des

connaissances acquises dans le passé.

ABSTRACT

The analysis of cellular structures in microscopy images is one of the most important tasks in various biological studies and disease diagnosis. In particular, segmentation and classification are two essential tasks that are done regularly in practice. However, the increasing heterogeneity, throughput and complexity of the data generated by modern microscopes have introduced various challenges for automatic computer algorithms. This is undoubtedly true for deep learning-based approaches. First, the manual creation of sufficient ground truth for training data in automated classification or especially segmentation is challenging since it requires a big commitment from domain experts. Then, given the limited labeled data samples for training, deep learning models often have low generalization ability. Furthermore, the high model complexity in terms of computing requirements and training process impacts the adoption of deep learning solutions. Biologists or pathologists find it difficult to choose one existing deep learning model over the others for their application having specific specimen preparation and image acquisition settings.

The main objective of this thesis is to develop novel deep learning architectures and techniques to improve the generalization and applicability of deep learning solutions for the segmentation and classification of microscopy images across various experimental conditions, such as microscope types, imaging conditions, sample preparation processes and cell categories. The first study investigates whether it is possible to eliminate the dependence of the deep learning model on the manual labeling process for the segmentation of nuclei in histopathology images. We developed an unsupervised DL segmentation model that is domain-adaptable. Our approach focused on exploiting prior knowledge of physical parameters in the image acquisition process and combining deep learning architecture with classical algorithms. We also design a data synthesis pipeline to generate augmented datasets containing examples resembling each of the target images and having their corresponding masks. The results which were recorded on three public datasets of histopathological images demonstrate that our method can outperform other unsupervised segmentation approaches and be comparable with supervised DL models.

In the subsequent studies, we focused on addressing the question of whether it is possible to develop deep learning-based classifiers that have good generalization across different applications given very limited training data. We first developed a lightweight CNN model to classify microscopy images in multiple domains simultaneously. The target data include images that captured different levels of cellular structures, from tissue to cells and cellular organelles. In addition, we elaborate an effective optimization procedure allowing the proposed network to outperform state-of-the-art methods without requiring the adaptation of domain-specific parameters. Interestingly, our proposed model is robust against limited available training data. With low complexity and wide applicability, the approach becomes more appealing for deployment with limited hardware resources in laboratory settings.

On the foundation of the above generic model, we then explored the possibility to improve further the accuracy of a specific application, particularly the recognition of subcellular organelles. We initially built a novel deep learning architecture that combines DL-based feature extraction with multi-resolution analysis. Experiments show that it can significantly improve the classification performance compared with state-of-the-art deep learning models on the same datasets of microscopic fluorescent images. Following this model, we extended the research with the development of another similar compact DL network. Unlike the previous model in which the feature extractor layers were pre-trained on the popular ImageNet dataset, we train the new model from scratch, without using any external images. We also formulated a regularizer to integrate into the optimization function. By experiments, our proposed model has surpassed state-of-the-art methods in similar experimental conditions and more importantly, it can generalize well even with very limited labeled data. In particular, the model requires about four times less than other approaches to achieve a similar level of accuracy. We believe the developed methods in this project will contribute to significantly decreasing the burden of generating ground-truth annotation for training data and increasing the analysis efficiency of pathologists and biomedical scientists. Further research should work on designing an end-to-end pipeline that can both perform segmentation and classification of the objects of interest. A method of continual learning should also be studied to help the learning agent be ready to scale to the new tasks based on the knowledge learned from the past.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ACRONYMS

ADC	Average Dice Coefficient
AI	Artificial Intelligence
AJI	Aggregated Jaccard Index
ANN	Artificial Neural Network
AUC	Area Under Curve
BN	Batch Normalization
CNN	Convolutional Neural Network
CHO	Chinese Hamster Ovary
DA	Domain Adaptation
DEC	Deep Embedded Clustering
DNN	Deep Neural Network
DCNN	Deep Convolutional Neural Network
DNA	Deoxyribonucleic Acid
FCN	Fully Convolution Neural Network
FPR	False Positive Rate
GAN	Generative Adversarial Networks
HSV	Hue-Saturation-Value
H&E	Hematoxylin and Eosin
MDL	Multi-Domain Learning
MoNuSeg	Multi-Organ Nucleus Segmentation
MMD	Maximum Mean Discrepancy
RGB	Red-Green-Blue
RNA	Ribonucleic Acid
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
ROC	Receiver Operating Curve
SAE	Stacked Autoencoder
SDL	Single Domain Learning
SIFT	Scale Invariant Feature Transform
SVM	Support Vector Machine
TCGA	The Cancer Genome Atlas
TNBC	Triple-Negative Breast Cancer
TPR	True Positive Rate

VAE	Variational Autoencoder
WSI	Whole Slide Image

LIST OF APPENDICES

CHAPTER 1 INTRODUCTION

Cells are the fundamental biological units of living organisms and one of the most important targets of microscope image analysis in biological studies or disease diagnosis. Depending on a particular application, pathologists may want to evaluate the variation in dimension or shape of cells in a region of interest because these properties are highly indicative of the cell's physiological state [10]. In other applications, researchers may focus on investigating the distribution of certain proteins or sub-cellular organelles within a cell.

Although analysis of cellular information is essential, the capability of human experts is limited due to an increasingly large collection of microscopic imaging data, in which complex patterns exhibit a complicated relationship with diseases [4]. Besides, there is an increasing heterogeneity, throughput and complexity of the data generated by emerging modern devices [10]. Manual evaluation of pathology samples is also affected by both subjective expectations and experience from pathologists before obtaining experiment results. Thus, the development of an automated analysis method that enables accurate and reproducible cell quantification is indispensable [11].

Segmentation and classification of cell structures are two essential tasks in the practice of pathology, including both histology and cytology, and biological studies. Segmentation is the process of clustering pixels into salient image regions corresponding to objects of interest or parts of objects. This is done by assigning a label to every pixel in an image such that pixels within a region share similarities with respect to certain characteristics or features, such as intensity, color or texture. After segmentation, the location and boundary of objects in the image are determined whereas background regions containing irrelevant information are discarded. In microscopy image analysis, segmentation can help researchers to focus on the useful area with rich information from data [12]. For example, it is used to study the cellular morphology which is an indicative phenotypic feature for the physiological state of a cell [10]. The cellular contours extracted from the image are also used to separate between cells, which is required for analyzing intra-cellular processes or studying cell sociology. The output of the segmentation could be semantic-level or instance-level segmentation map. While semantic segmentation simply categorizes each image pixel as foreground or background, instance-level segmentation distinguishes different objects in the foreground. Low-quality segmentation certainly impacts the subsequent analysis and in practice, different segmentation algorithms often produce different results [13].

On the other hand, classification is the process of assigning one of the predefined labels to an image of a separated object. The separation could be done by either segmentation or detection algorithms. Unlike the aforementioned segmentation, the goal of detection algorithms is to recognize the location of a target object in an image without specifying exactly its contour. As the labels are context-dependent and user-defined, there exists a wide range of microscopy image classification tasks. Some of the examples include distinguishing different subcellular organelles or cells, grading cancer stages or identifying various diseases.

Microscopy image acquisition systems

As described in [14], there is a sequence of required preparation steps (Figure 1.1) so that a tissue sample on the slide could maintain structural features similar to those of a natural living state. In the fixation step, sample tissue is placed into chemical solutions to preserve cell and tissue structure by cross-linking proteins and deactivating degradative enzymes. As a large composition of tissue is water, it is passed through a series of alcohol solutions with increasing concentration to remove water (dehydration) and then treated with organic solvents which make tissue becomes transparent (clearing). In the next infiltration step, the tissue is usually put into melted paraffin for a certain time to be completely infiltrated with this substance. After that, the paraffin-infiltrated tissue is placed into a small mold covered with melted paraffin to become hardened and form a paraffin block containing the tissue (embedding). It is necessary for sectioning that block of tissue into thin sections by a microtome, for imaging by trans-illumination in optical microscopy. Paraffin sections are then mounted on glass slides (optical microscopy) or metal grids (electron-microscopy) for staining and examination. As different tissue constituents may have similar optical densities, they should be stained for easy discrimination, usually with water-soluble stains (optical microscopy) or metal precipitation (transmission electron microscopy). For optical microscopy, this step requires removing the paraffin first, then rehydrating before staining. The most commonly used stains are hematoxylin and eosin (H& E). They produce bluish color for the acidic components of a cell (e.g, nucleus) and a pinkish tint for basic components (e.g, cytoplasmic). Finally, the tissue slide could be imaged with different microscope devices to generate microscope images. Unfortunately, there exists variation at different stages of the sample preparation since each stage has a range of setup parameters, causing large variability across different clinical laboratories and within the same lab over time [15, 16]. This directly results in a lack of consistency in microscope image quality and thus, analysis results.

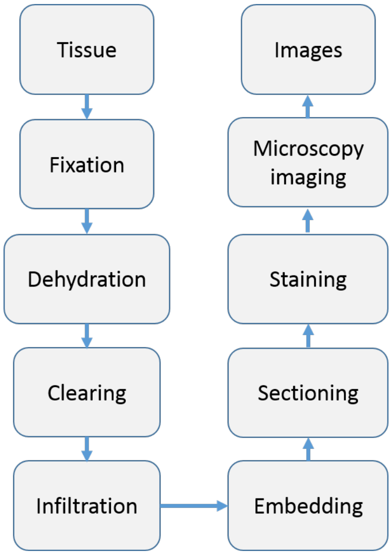


Figure 1.1 Sample preparation steps for microscopy imaging

Common types of microscopy devices and their images

Light microscopy is the dominant technology for research and diagnosis worldwide. It is based on the interaction of visible or ultraviolet light with tissue components to visualize images. Different types of light microscopes can be used to complement each other in diagnosis and research, including conventional bright-field microscopes and later developed specialized types like dark-field, fluorescence, phase-contrast, confocal scanning laser, differential interference contrast, and two-photon microscopes. The most important factors in microscopy imaging are magnification, resolution, and contrast [2]. Magnification enlarges the appearance of the specimen in the image. To make two objects distinguishable, magnification should be set to a certain level given an appropriate resolution, otherwise, additional details cannot be seen clearly. Light microscopes can magnify effectively up to 1,000 times the actual size of the specimen. Resolution refers to the minimum distance between two points to be distinguished as separate points. Typically, the light microscopes can not resolve detail finer than about $0.2\mu m$. Lastly, contrast is the difference in brightness between the light and dark regions of an image and could be enhanced by staining or labeling cell components. Table 1.1 summarizes different light microscopes and their example images [1, 2].

1.1 General research objectives

In practice, microscopy image introduces various challenges for designing automated segmentation and classification algorithms. Direct segmentation of cell or sub-cellular structures is very challenging when dealing with, for example, heterogeneous intensity, fading boundary, overlapping cells and background clutter [4]. Compared with medical radiology images which focus only on a few organs of easily predicted location, microscope images normally contain a vast amount of objects (cells and organelles) randomly located and/or surrounded by complex tissues [17]. Meanwhile, obtaining sufficient data ground truth for automated classification (image-wise labels) or segmentation (pixel-wise labels) is not feasible since it requires a big commitment from domain experts [18]. Moreover, large variability in different microscopy imaging devices, staining processes and cell types make it difficult to achieve generally optimal results. Besides, microscopic data become increasingly complicated and multi-dimensional (multi/hyper-spectra, multimodal microscope images, and ultra-large whole slide images). Considering these situations, applying any analysis approach alone could not produce satisfactory results even when tailoring to a specific problem. It seems to be impossible to converge to a robust, universal solution: there are almost as many methods as there exist cell analysis problems [4]. This prevents pathologists from adopting applicable solutions for their clinical setting with specific specimen preparation and image acquisition [18]. There is an increasing demand for a generic model which is applicable for a wide range of images generated with different pathology protocols [19, 20].

Deep learning models have achieved state-of-the-art performances in many medical image analysis applications. Its astonishing success is often dependent on the amount of annotated data examples and the complexity of deep learning models. For microscopy image analysis, these two requirements are the big hurdles for the generalization and applicability of deep learning-based methods when deployed in clinical settings. The overall goal of this research project is to improve the generalization and applicability of deep learning algorithms for the segmentation and classification of microscopy images across various experimental conditions in terms of various microscopes, imaging conditions, sample preparation processes and cell types. Firstly, the research will propose a segmentation algorithm that generalizes well on multiple domains of acquired microscopy images without the need for reconfiguration for every specific application. Secondly, the research will propose compact classification algorithms that have strong generalization capability given the condition of limited annotated training data. In general, we will focus on developing deep learning based methods that generalize well with very limited data while still maintaining their computational efficiency.

1.2 Manuscript Overview

The remainder of this thesis is organized as follows:

- Chapter 2 reviews related works on segmentation and classification algorithms of microscopy images. Next, basic concepts and well-established standard methods aiming at improving the generalization of deep learning models will be summarized.

- Chapter 3 first describes the current trends and challenges that justify the motivations of this research project. Then, the research objectives to achieve the overall goal are established.

- Chapter 4 details our method for designing a domain-adaptable model for nuclei segmentation in histopathological images using unsupervised learning. Its main goal is to eliminate the dependence on the expensive data annotation process in training a nuclei segmentation model.

- Chapter 5 presents the second contribution which proposes a multi-domain learning CNN model for microscopy image classification. This work aims at designing a generic model that can apply to multiple imaging domains, instead of having multiple per-domain models.

- Unlike the generic model in the previous chapter, we next present our contribution that focuses on a specific application of fluorescent microscopy image analysis, for automatic recognition of sub-cellular organelles, in chapter 6. In this work, we design a lightweight architecture that helps to improve the performance and generalization of the deep learning model.

- In connection to the above study, chapter 7 provides additional research that demonstrates the benefit of using compact deep learning design on the cellular organelles classification with very limited annotation.

- In chapter 8, we discuss the advances and limitations of our works in relation to previous studies.

- The final chapter 9 summarizes the major contributions and provides perspective on some potential research directions in the future.

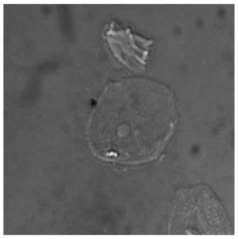
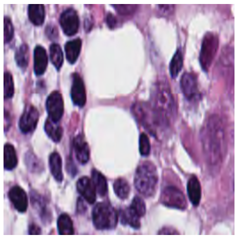
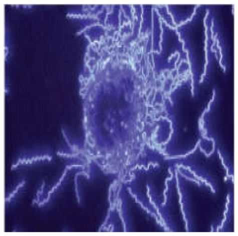
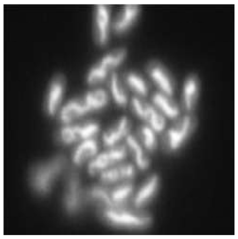
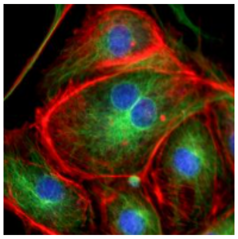
Microscope type	Example Image
Bright-field without staining: Bright-field microscope is widely used as the standard type, in which light passes directly through the specimen. Without staining, the image has little contrast if the cell is not naturally pigmented or artificially stained.	
Bright-field with staining: The contrast of an acquired image is enhanced by staining the cell sample with specific dyes, in which process cells are fixed (preserved) and thereby are killed.	
Dark-field: Instead of using directly transmitted light, only light scattered off the specimen is used to produce the high contrast image without the use of stains.	
Confocal: Uses a laser to scan multiple z-planes successively to generate a sequence of high resolution images at various depths that allows reconstructing a 3D image of the thick specimens. Each image slice is sharp compared with a blurry standard image because out-of-focus light is excluded.	
Fluorescence: Fluorescent substances absorb ultraviolet radiation and emit visible light. Some cells inherently have molecules that fluoresce on their own, or molecules could be labeled by fluorophores.	

Table 1.1 Light microscopes use visible or ultraviolet light to produce an image [1,2]. Royalty-free Images.

CHAPTER 2 LITERATURE REVIEW

Microscopy is one of the most important technologies for research and diagnosis worldwide as it appears in around 90% of publications in life sciences according to a recent survey [21]. We start this chapter by reviewing state-of-the-art segmentation and classification techniques for microscopy images. For each task, we will first summarize the principles of conventional non-learning-based methods and then focus on deep learning algorithms that have recently attracted considerable attention in the literature. In the last section of this chapter, we summarize the efficient techniques for improving the generalization of deep learning models and then review the application of these techniques in different approaches for microscopy image analysis.

2.1 Segmentation

Segmentation, which is also formulated as pixel-wise classification, is the process of separating every cell or sub-cellular compartment to allow measurement of morphological characteristics such as intensity, size, shape or distribution. Although segmentation of objects in microscopy images is regular, it is not a simple task. Manual segmentation is expensive, as it is time-consuming and requires the expertise of experienced image analysts [22, 23]. Therefore, computer algorithms that can produce high-throughput analysis at the expert level are much-needed [21].

General criteria for a complete segmentation are [24]:

1. No pixel is without a region assignment.
2. Each pixel has a unique region assignment.
3. Pixels in each region are connected.
4. Each region satisfies a given uniform predicate.
5. Any merged pair of different regions is non-uniform.

2.1.1 Conventional approaches

Although automated image segmentation for cell analysis is a generally challenging problem and has been applied in microscope image analysis for many decades, there are essentially just a few principal methods. They are typically developed for other fields of computer vision before being adopted for cell segmentation. Figure 2.2 summarizes the common conventional cell segmentation approaches after screening literature published during period from 2000 to 2012 [10]. Other specific methods include dynamic programming, graph cuts, active masks, support vector machines, tensor voting schemes, neural networks and Markov random fields.

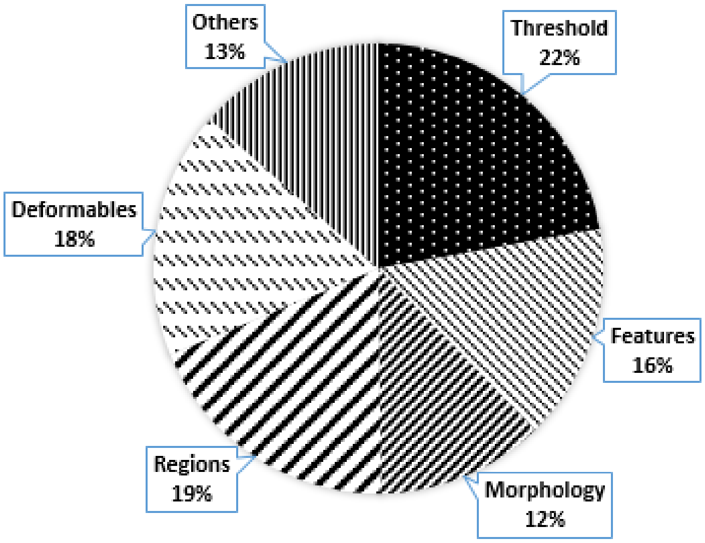


Figure 2.1 Works of literature on conventional cell segmentation methods in 143 journal papers from 2000 to 2012.

Intensity thresholding

This is one of the most popular approaches due to its simple computation and high success rate against solid objects with closed, connected boundaries on a contrasting background [13]. The thresholding operation is described as:

$$G(x,y) = \begin{cases} B & \text{if I(x,y) < T} \\ F & \text{otherwise} \end{cases} \tag{2.1}$$

Where $G(x, y)$ is the thresholded image which is divided into background B and foreground F corresponding to objects, $I(x, y)$ is the original image whose gray-level or another extracted feature (that has been converted to gray-level) is compared against a specific threshold value T. The boundary is then determined by checking the set of foreground points to have at least one neighbor outside the object. Notice that the threshold value could be applied globally (fixed thresholding) or locally (adaptive thresholding) and for automated threshold selection, statistical analysis of intensity histogram is usually used [10].

Feature detection

This kind of segmentation method is based on extracted features from images by linear image filters instead of using intensity level directly [10]. For images with high cell densities and intensity variation, this approach could be more effective than intensity thresholding. For example, [25] presented a method for cell nuclei detection and segmentation, based on a multi-scale Laplacian of Gaussian (LoG) filter which was considered as a generic blob detector. It was demonstrated to offer advantages such as accuracy improvement, computational efficiency and robustness to variations. Alternatively, first derivative operators (edge detection) or second derivative operators (ridge detection) are used for locating directly the boundaries by identifying the edge pixels [13]. The edge detection or gradient-based methods look for edges by detecting the maximum in the first derivative of an image. On the other hand, ridge detection or Laplacian-based methods search for zero-crossing in the second derivative of the image. In practice, the edge points can not form completely closed connected boundaries and therefore the subsequent step of edge linking is required to associate nearby edge points.

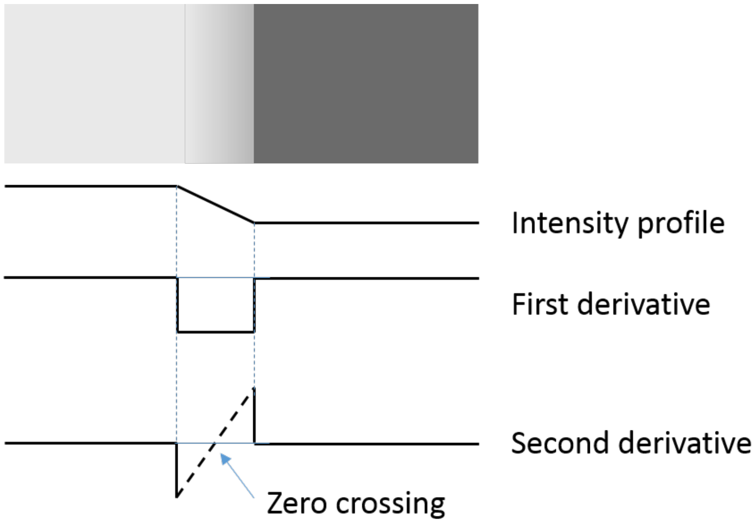


Figure 2.2 An edge and its first and second derivatives [3].

Morphological filtering

Morphological filtering uses non-linear operations related to geometrical and topological properties of objects in images, such as erosion, dilation, opening and closing. Its principle is to probe an image with a small structuring element and test whether the structure element fits in each corresponding neighborhood of pixels. There are basically two categories of morphological techniques: binary morphology could be used as a post-processing step to improve the initial segmentation results from thresholding by for instance separating touching objects and filling internal holes [13]; gray-scale morphology is usually used as pre-processing image enhancement in order to eliminate certain image structures before segmentation [10]. Complicated compound filters could be constructed by successive application of different morphological operations. For example, in [26], a multi-scale decomposition method for cell segmentation was proposed, consisting of top-down erosion and bottom-up dilation procedures. Morphological operators are generated by using scalable templates at multiple levels. First, distinct markers for each cell are identified in a top-down procedure by applying erosion iteratively on the noise filtered binary to erode regions down to markers until a steady-state is reached. Then, the bottom-up dilation procedure reconstructs the original shapes from markers while maintaining their separating status.

Region based segmentation

In contrast to the intensity thresholding method which groups pixels without considering their spatial location in the image, the region accumulation starts from a set of predefined seed pixels and iteratively adds connected pixels exclusively to each region [24]. The simplest implementation of this method is the region growing, which employs a uniformity predicate to control the adding of a pixel to a region, meaning that the addition of a pixel must preserve the uniformity of the growing region. In general, it is very sensitive to choose uniformity threshold and different selection of hyper-parameters such as seeds, types of pixel connectivity or routes of scanning an image could result in different segmentation maps. For example, the number of preselected seeds may not match with the required number of regions or two seeds stay within a potentially uniform region still results in two distinct regions [27]. An improved version is the hierarchical split-and-merge algorithm which either (top-down) iteratively splits the initial entire image into sub-regions or (bottom-up) merges adjacent regions until all regions become uniform or the desired number of regions is reached. The most popular approach is watershed transform, which views any gray-scale image as a topographic surface where high intensity or gradient magnitude represents hills and low value denotes valleys. The philosophy behind can be intuitively described via the process of gradually filling colored water (labels) to every isolated valley (local minima) [28]. The segmentation edges correspond to barriers that prevent colored water of nearby valleys from merging when its raising level is higher than local peaks. The most challenging problem with watershed transform is the resulted over-segmentation due to noise or any irregularities in the image and thus further processing is usually required. A practical widely used solution for cell segmentation is the marker-controlled watershed which provides criteria for merging regions to solve the over-segmentation [29]. In this approach, before applying the watershed algorithm, valley points to be merged are specified and no barriers will be created between them if their flooded areas happen to merge [30]. As such, regional minimum produced by noise or minor structure in an image will not become a marker or center of the growing region, and thus significantly amend the over-segmentation. In order to label the markers, automated cell detection methods are preferred over manual marker labeling, especially in the case of large-scale images.

Deformable model fitting

This method represents the mechanism for conditioning deformable models such as a curve (for 2D image) or a surface (for 3D image) to fit an object boundary. In order to constrain the evolvement of these models over space and time, energy function must be formulated and minimized. For example, in the active contour model (also referred to as snake), the energy function is a linear combination of three terms:

$$E_{snake}(v) = E_{int}(v) + E_{img}(v) + E_{con}(v) \tag{2.2}$$

Here a simple elastic snake consists of a set of points v_i. The internal elastic energy term $E_{int}(v)$ is composed of the continuity and the smoothness of the contour to control the deformations made to the snake. Next, the image energy $E_{img}(v)$ is calculated as some function of different image features like lines, edges, terminations. Finally, the constraint energy $E_{con}(v)$ allows the user to interactively guide the snakes near the desired features. The combination of E_{img} and E_{con} is regarded as the external edge-based energy which controls the fitting of the contour onto the image object [31]. In general, there are two main implementation approaches for deformable models depending on practical applications [29]. Firstly, the geodesic or level set model represents a contour implicitly as the zero level set of a high-dimensional function with one dimension higher than the image to be segmented. As there are many cell objects within a microscope image, energy functions of a multi-level set are usually adopted. Secondly, the parametric model represents a contour or a surface explicitly by a continuous parameter $v(u) = (x(u), y(u))$,with $u \in R | 0 \leq u \leq 1$. By minimizing the energy function, the initialized contour evolves toward desired features while still it is kept smooth. In practice, the parametric models require lower computing and time complexity than corresponding geodesic models. To cope with cell touching, an additional repulsive term is incorporated into the energy formula to prevent adjacent contours from merging and help to separate touching cells. For an image with overlapping cells, using shape prior constraint is widely applied [32, 33]. In all cases, designing relevant energy terms is particularly necessary to avoid erroneous segmentation.

2.1.2 Deep learning approaches

Deep neural network or deep learning is an advanced technology applied for image pattern recognition. Especially since 2012 when AlexNet model [34] won ImageNet challenge [35], Convolutional Neural Networks (CNNs) have become the most dominant trend for research in image classification. The number of papers describing applications of deep learning to medical image analysis, including microscope image analysis, increases rapidly starting from 2015 [21, 22]. Recently, even in the area of clinical diagnosis, where people have the conventional belief that machines cannot deliver human competitiveness, pattern recognition based on deep learning demonstrates outstanding performance [36]. Deep convolution neural networks such as GoogleNet [37] and ResNet [38] have been applied to some applications in medicine and achieve excellent recognition accuracy comparable to human performance. Unlike methods based on hand-crafted feature extraction, deep learning algorithms can extract optimal discriminant feature representations directly from raw image data by optimizing a cost function defined for a specific task like classification or segmentation. However, training a deep learning model generally suffer from serious overfitting problem or low generalization capability due to commonly limited small datasets with just dozens to hundreds of images per class. In this case, a practical method that has been reported in many publications is transfer learning [39]. In particular, the feature extraction module of a network previously trained on another large dataset is used to extract useful features of objects in microscope images. In practice, previously trained networks can be fine-tuned to produce satisfactory results on new datasets of similar complexity and of limited training data. So far, CNN and its variants are the most popular architectures for a wide range of image analysis applications, including microscope image segmentation [4, 22]. Some other popular architectures are Fully Convolution Neural Network (FCN), Recurrent Neural Network (RNN) and Stack Autoencoder (SAE). Figures 2.3 and 2.4 show the distribution of 103 deep learning papers on microscope image analysis (including detection, segmentation and classification) according to network architectures and different types of microscope images [4].

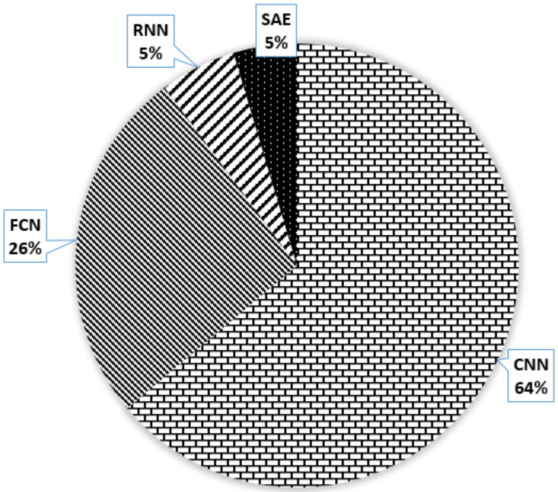


Figure 2.3 Distribution of 103 deep learning papers on microscope image analysis according to network architectures in a recent survey [4].

Convolution Neural Network (CNN)

Similar to a traditional neural network, a convolution neural network comprises multiple hidden layers between input and output layers. Each hidden layer is composed of neurons or units, each one has a set of weight-bias parameters and an activation function. For a network of L hidden layers, each layer k^{th} where $(1 \leq k \leq L)$ is associated with a set of connection weights $W^{(k)}$ and biases $b^{(k)}$, the layer pre-activation is (Courville, 2019):

$$a^{(k)}(x) = W^{(k)} * h^{(k-1)}(x) + b^{(k)}$$

Where activation $h^{(k)}(x)$ is calculated via a hidden layer activation function g:

$$h^{(k)}(x) = g(a^{(k)}(x))$$

The common functions g are sigmoid(), tanh() or ReLU(). Then, the network output is computed via an output layer activation function o:

$$f(x) = h^{(L+1)} = o(a^{(L+1)}(x))$$

For binary classification, activation function could be the same as in hidden layers, i.e sigmoid() or tanh(), whereas for multiclass classification, a softmax function is generally chosen:

$$o(a) = \text{softmax}(a) = \left[\frac{\exp(a_1)}{\sum_c \exp(a_c)}, ..., \frac{\exp(a_c)}{\sum_c \exp(a_c)}\right]$$

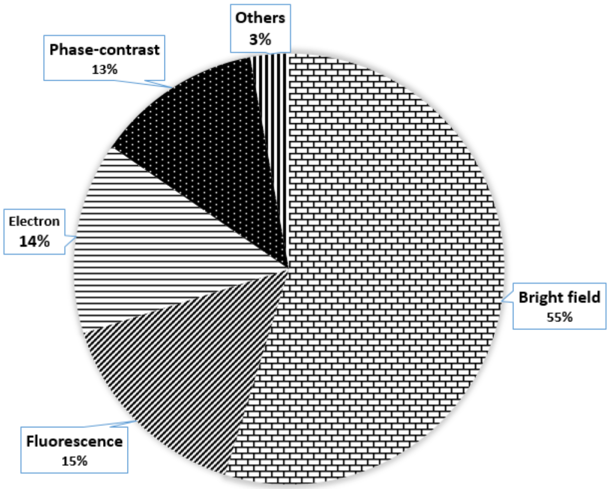


Figure 2.4 Distribution of 103 deep learning papers on microscope image analysis according to different types of microscope images in a recent survey [4].

Compared with the traditional neural networks, a convolution neural network could deal with very high-dimensional variable inputs and exploit the spatial topology of pixels based on the mathematic operation of convolution [40]. Firstly, each hidden unit only covers a small sub-region or a patch of the input image instead of being connected to all pixels as in a fully connected hidden layer which has an unmanageable number of parameters for a large input image. This avoids the computing of expensive activation functions for hidden units. Secondly, a convolution neural network shares the matrix of parameters across units of the same feature map. This not only reduces the number of parameters that need to be stored but also helps to produce features that are equivariant to translation. In other words, a translation of input data results in the same extracted features having an equivalent translation. Finally, it could introduce invariance to small local translations and reduce the number of hidden units in each layer by pooling or subsampling operations. Figure 2.5 shows a simple CNN architecture where convolution and pooling layer alternates each other.

To segment an image into different components or classes, CNN is typically used to classify each pixel individually by sliding a window on input images and generating probability maps. Features in a patch around every pixel are extracted by non-linear filters of the network and segmentation labels are produced separately. Authors in [41] introduced an end-to-end framework based on the convolutional neural network which is proposed to segment raw pixels of microscopic images into five different categories including cell wall, cytoplasm, nucleus membrane, nucleus and outside medium. The label for each pixel in the input image is produced by a convolution network composed of interleaving 3 convolution and 2

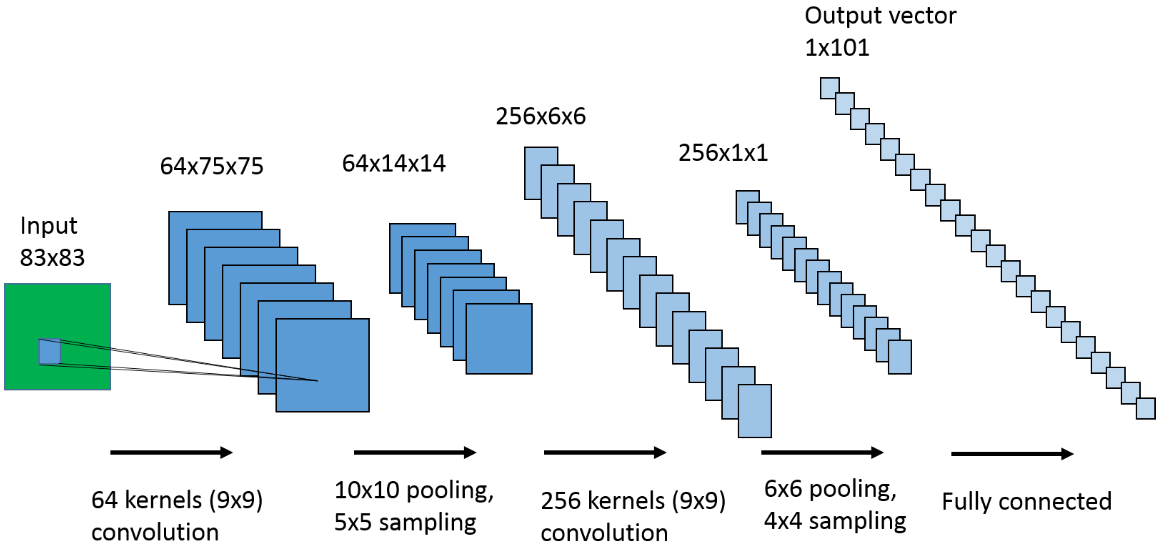


Figure 2.5 A simple convolution neural network [5].

subsampling layers before a fully connected layer, applied on a 40 × 40 pixel window. Thus, the full network which produces the full segmentation map could be viewed as multiple replicas of such network sliding on input image with a step of four pixels. As each label is produced independently with the labels of neighboring pixels and thus introducing local inconsistency, a set of energy-based constraints is implemented to clean up the output of the convolutional network. In [42], a similar CNN architecture is proposed to be used as a binary classifier for segmentation of biological neuron membranes in transmitted electron microscopy (TEM) images. After training, the classifier segments a test image by using the softmax function to compute the probability of each pixel being one of two possible classes which are membrane and non-membrane. Indeed, the unbalanced testing set causes a severe over-estimation of membrane probability in the output. Thus, a monotone cubic polynomial is approximated to transform the network outputs and calibrate the final results. Data augmentation and non-uniform sampling techniques are additionally applied to input training images to extract invariant features and exploit data at multiple resolutions. Finally, averaging calibrated outputs of multiple network architectures was demonstrated to improve significantly the performance score against single networks. As an extension, [43] proposes to use a multi-scale convolutional network that is composed of multiple parallel networks to process input images at different scales for segmentation of cervical cytoplasm and nuclei. However, this resulting coarse segmentation by the network alone is not visually satisfying and thus a fine-tuning step by graph partitioning and problem-specific manual post-processing by constraining the result under prior knowledge are necessary. Later, CNNs with problem-specific optimized hyperparameters are designed for different segmentation applications such as breast cancer data [44], skeletal muscle images [45], brain electron microscopy [46].

Fully Convolutional Network (FCN)

A major drawback of using CNN classifiers for pixel-wise segmentation is its efficiency, as input patches from neighboring pixels overlap and thus convolution operations are computed redundantly [22]. Fully convolutional networks overcome this challenge by replacing the fully connected layers with convolution layers and they produce an output map instead of a single label for each pixel. However, due to subsampling layers, the output dimensions or resolution are reduced. Thus [47] adds bilinear upsampling layers to obtain corresponding size output and uses skip layer fusion that combines multiple layers of high resolution features with upsampled output to refine spatial details. This strategy also helps to mitigate the trade-off between localization accuracy and patch context information in conventional CNN approach [6]: larger patches provide more context but decrease localization accuracy. Further, [48] designs a deep contextual network to upsample and then sum up feature maps at multiple levels of the downsampling path. Some auxiliary classifiers are injected into the network to provide a regularization term for the loss function. This technique is considered to deal with the vanishing gradient problem for deep networks, to reduce overfitting and improve the discriminative capability of extracted features. Noticeably, [6] extends the concept of FCN by proposing the well-known U-net architecture which consists of an expansion path that is roughly symmetric to the contraction path. U-net upgrades the learnable upsampling part in combination with skip-connection such that context information from contracting layers could be concatenated with expanding layers of corresponding resolution level (Figure 2.6). Besides, several techniques are also applied to enhance the performance and robustness of the network for the segmentation of microscope images. Firstly, data augmentation of training images including shift, rotation, gray variation and especially elastic deformations enable the network to learn invariant features with very limited annotated images. The addition of drop-out layers also provides an additional implicit data augmentation effect. Then, a weighted loss function is proposed to deal with touching cells, where background labels separating touching cells in the ground-truth are assigned more important weights. Until recently, U-net-based architectures have been widely used to segment various cells and tissues in different types of microscope images such as neural membranes in TEM images, HeLa cells in differential interference contrast images, glioblastoma-astrocytoma cells in phase-contrast microscope images [4].

The FCN structure could also be used to extend object detection networks, such as Faster R-

CNN [49], to develop instance segmentation model Mask R-CNN [50]. This architecture adds the FCN layers as a branch for object mask prediction in parallel with the existing branch for bounding-box object detection of the Faster R-CNN. This design allows the separation between different entities of the same segmentation class. The Mask R-CNN architecture has been adopted recently for nuclei segmentation [51, 52].

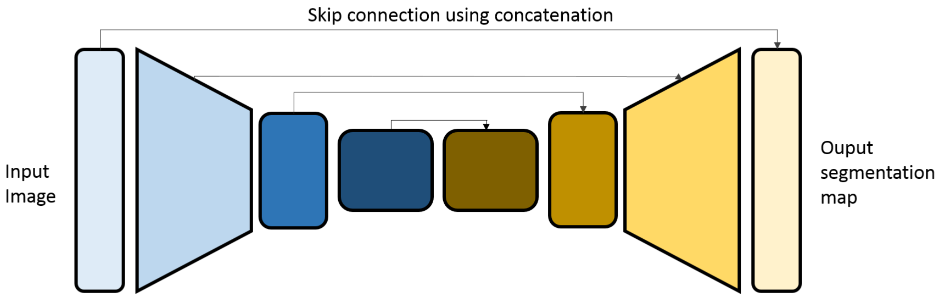


Figure 2.6 U-net architecture [6].

Recurrent Neural Network (RNN)

Figure 2.7 shows a simple RNN model which is a class of neural networks that generate output depending on the persistent hidden state and previous outputs. As it forms a directed graph between nodes along a temporal sequence, the RNN model is typically used for natural language modeling or sequence processing. Formally, given a sequence $x_1, x_2, ..., x_T$ the model produces the conditional probability $p(y|x_1, x_2, ..., x_T)$ for classification purpose. The vanilla architecture consists of a hidden layer that outputs a non-linear mapping h_t at a time t from input x_t and previous state $h^{(t-1)}$:

$$h_{(t)} = \tanh(Ux^{(t)} + Wh^{(t-1)} + b)$$

Where U, W, b are shared weights (for input and hidden state) and biases over time. Moreover, to generate a prediction, fully connected layers are typically added and computed by a softmax function:

$$o^{(t)} = Vh^{(t)} + c$$

$$\hat{y} = \text{softmax}(o^{(t)})$$

To train the RNN model, the loss gradient should be obtained by applying the chain rule on

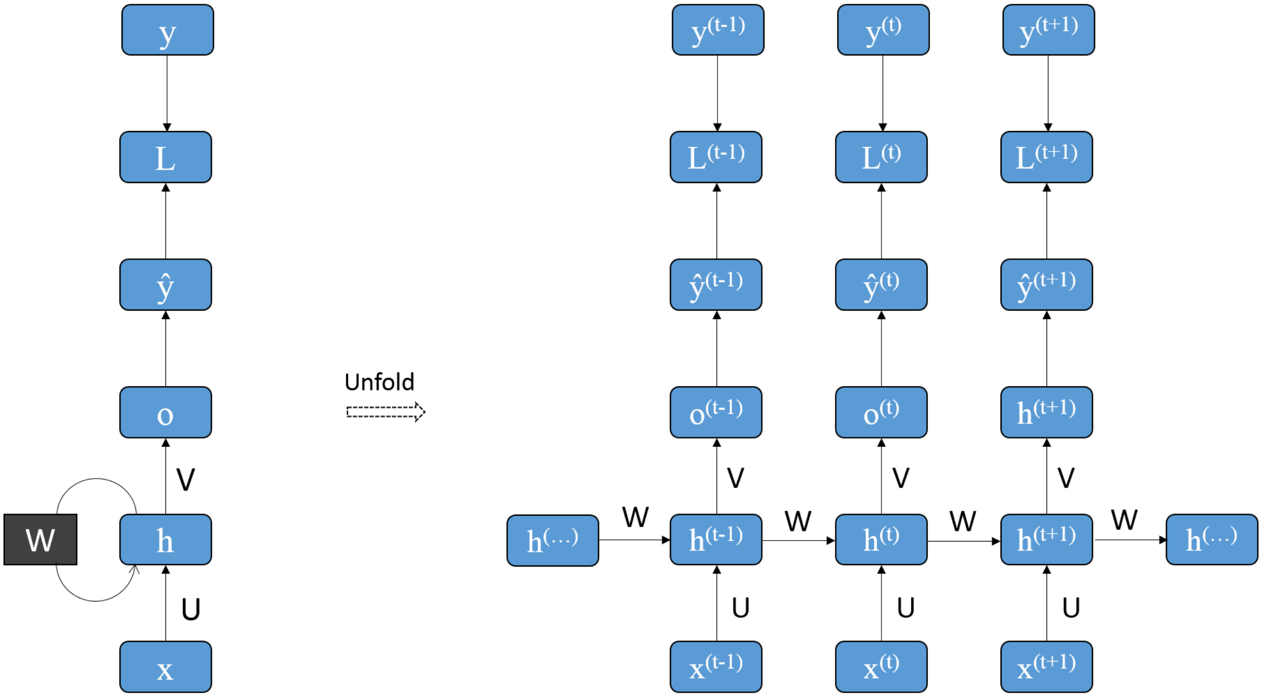


Figure 2.7 Vanilla Recurrent Neural Network architecture [7].

the unrolled graph which could lead to the problems of vanishing or exploding gradients for very long sequences. Therefore, a more popular architecture named Long Short Term Memory (LSTM) based on a specialized memory unit was proposed for learning long-term dependencies [53]. Its architecture is composed of gates that can control the flow of information in and out of a cell where information could be maintained over a long period, which is not the case for traditional RNNs. An improved version named Gated Recurrent Units (GRU) network is recently introduced with the same working principle but requires much fewer parameters and computations [54]. Recently RNNs are increasingly applied for biomedical image segmentation problems and they produce promising results [22]. For example, two-dimensional LSTM which connects hidden LSTM units in four directions is applicable for image segmentation. Each LSTM unit processes a pixel at a time while receiving outputs from preceding units as inputs such that information of other pixels in the image are recursively gathered. [8] modifies the topology of conventional 2D-LSTM into PyraMiD LSTM (Figure 2.8(c))which is optimal for GPU parallelization and computation, especially for 3D data. In particular, a 2D-LSTM adds the pixel-wise outputs of 4 LSTMs, each scanning the image diagonally from top-left, top-right, bottom-left, bottom-right. Figure 2.8(a) shows example for one direction from top-left. In this topology, although pixels on a simplex (dashed-line) could be processed in parallel, the number of such pixels on each dashed-line is not uniform making it difficult for parallelization in practice. Therefore, this topology is turned 45° (Figure 2.8(b)) and finally added extra connection to fill generated gaps. Some hyper-parameters like larger input filters are adjusted accordingly to the new computation fashion. As the experimental results show, this model achieved comparative results for segmentation of neuronal membrane electron microscopy images and the best result for brain magnetic resonance images.

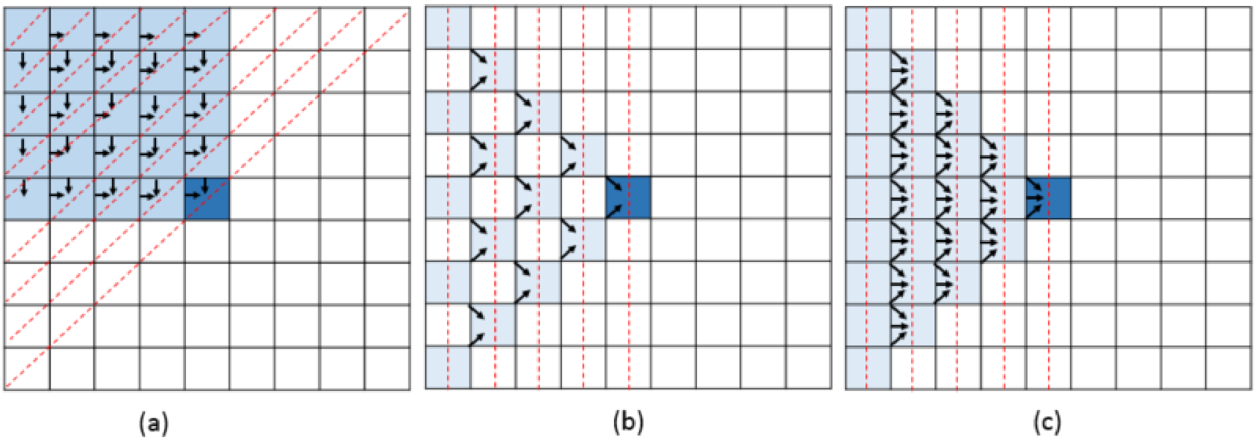


Figure 2.8 Improvement of 2D-LSTM topology [8].

Another interesting approach proposed in [55] divided the input image into a grid of patches and applied the structured regression to produce a prediction mask for each patch instead of a single label for each pixel. In addition, the global context information is obtained for each patch by the fact that activations from four entire-image-scanning following four diagonal directions are concatenated at each patch processing. This model is much more efficient and requires much lower parameters than LSTM based models while producing state-of-the-art accuracy for muscle perimysium microscope image segmentation.

Stacked Autoencoder (SAE)

SAE is a type of unsupervised model that leverages the availability of unannotated data for learning their meaningful features through adding a data-dependent regularizer to training, such as minimizing reconstruction loss [56]. SAE consists of auto-encoders placed on top of each other such that latent code (sparse features) of the lower autoencoder is fed as input to the higher autoencoder (Figure 2.9). Each autoencoder is composed of an encoder layer and a decoder layer, being trained to reproduce its input through a hidden layer. This hidden layer outputs a latent code which is a compression of input if the hidden layer is smaller than the input layer:

$$h = \sigma(Wx + b)$$

where h is the hidden layer activation or the latent code. W and b are the weight matrix and the bias of the hidden layer, respectively, while x is the input vector.

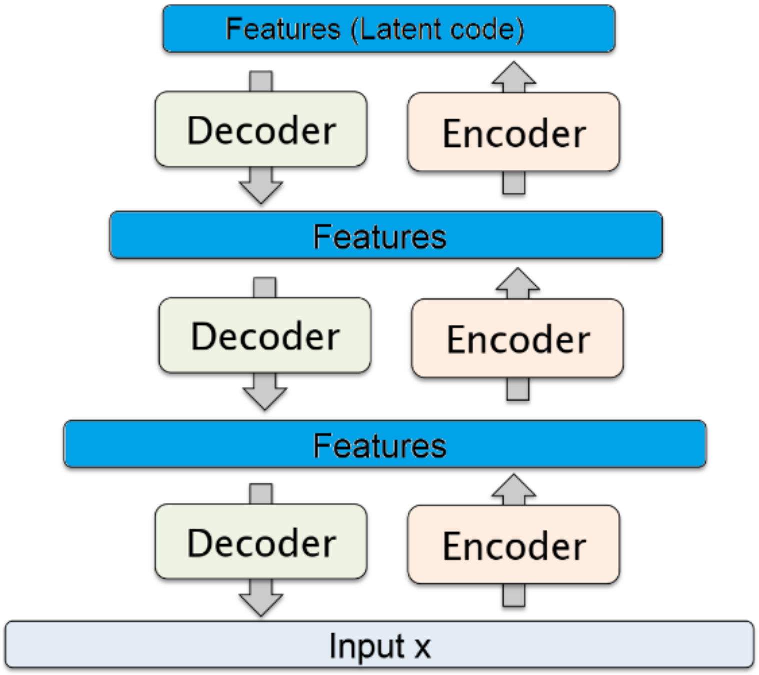


Figure 2.9 An edge and its first and second derivatives [3].

One popular solution named Stacked Denoising Autoencoder [57] enhances the robustness against noise by training the model such that it reconstructs the clean input from a noisy version, such as an added salt-and-pepper noise. This concept is applied in a model proposed in [58] for nucleus segmentation in histopathological images. It takes the noisy gradient patch around each pre-detected nucleus as input and reconstructs toward the annotated nuclear boundary of each cell. As a result, the experiments show that the trained model could remove fake edges and correct broken edges. However, in general, the major drawback of SAE networks is the fact that it usually requires individual layer-wise training or non-end-to-end fashion and then a fine-tuning process to achieve reasonable results. Although other recently introduced unsupervised architectures, such as Variational Autoencoder (VAE) and Generative adversarial network (GAN), have improved significantly the performance on natural images, there are few peer-reviewed papers for microscopy images [22]. One interesting recent work proposes to use a GAN-based framework for segmenting nuclei in different organs and produce better performance than conventional approaches [59]. However, a major challenge of the generative adversarial network is that it is notoriously difficult to train and its performance is unstable.

2.2 Classification

In microscopy image analysis, the classification task commonly involves recognizing different types of cells, subcellular organelles or identifying stages of a certain disease, for e.g cancer [12]. Unlike segmentation which outputs pixel-wise labels, microscopy image classification is the task of assigning a single label (for instance, disease or not) to an entire image. Since the last decade, researchers have been increasingly using deep learning-based classifiers in many applications as they usually offer higher accuracy and require less engineering time than conventional algorithms.

2.2.1 Conventional methods

Essentially, conventional classification algorithms rely on the extraction and selection of local characteristics in images such as points, edges or intensity distribution in neighborhood regions. These features have been manually and specifically designed in conventional hand-crafted methods for the last some decades. Well known handcrafted feature extraction approaches for microscopy images including Local Binary Pattern (LBP) [60], Zernike moment [61], Haralick texture features [62]. Popular efficient classifiers to discriminate extracted features are Support Vector Machine (SVM) [63] and Artificial Neural Network (ANN) [64]. In [65], to classify ten types of HeLa cells, a set of 174 features, including morphological features, Haralick texture features and Zernike moments in combination with wavelet-based filtering technique are selected as input for a majority-voting ensemble classifier made up of either neural network or SVM. Authors in [62] also try to design optimal feature sets with morphological, Haralick texture and Zernike moments features but instead of using wavelet-based features, they compute features in each decomposed sub-space. The local decisions in each sub-band are then combined into a global decision by a step of weighting fusion. They end up using 26 Haralick texture features and a neural network classifier for each sub-band. As usual, the most challenging problem is the transferability of such feature sets to their related datasets. Engineering the optimal features is time-consuming, error-prone and especially, default parameter values are not necessarily relevant for new data other than the domain for which they were crafted [66].

2.2.2 Deep learning methods

Up to date, there is a dominant trend of using deep learning in microscope image classification due to their obvious performance over traditional methods. Indeed, most methodologies are still developed for natural image analysis and proposed works in the microscope image domain attempt to adapt existing architectures to the specific problem at hand [22]. As mentioned above, two common applications of deep learning-based classifiers include cellular or subcellular classification and disease diagnosis [12]. Some studies have shown that deep learning methods were highly accurate at distinguishing different types of cells [67] or stages of cells differentiation [68, 69]. To assist pathologists in diagnosing blood-related diseases, researchers have developed classifiers to identify white blood cells [70–72] or recognize sickle cells from normal red blood cells [73]. Other works aim at detection of cancerous samples, such as for lung [74], breast [75] or colon cancer [76].

Among deep learning approaches, the use of convolution neural networks is becoming more and more popular for image classification. In practice, a CNN network could be used as a direct classifier to output a prediction or its layers are used as feature extractors being integrated into other classifiers. Normally, supervised learning-based networks outperform unsupervised ones and have been demonstrated as a powerful tool for microscope image classification [4]. For example, the authors in [77] utilize a simple LeNet-like CNN architecture to design an automatic framework for the classification of segmented human epithelial-2 cell images. Various hyper-parameters related to training an effective network are discussed and then the importance of data augmentation, such as rotation, is proved by analyzing its impact separately. Moreover, as a common strategy, fine-tuning a model pre-trained on a larger related dataset could significantly produce better results. Aiming at a more complex model for HeLa cells recognition, authors of [78] develop a multi-scale convolution neural network, consisting of 22 convolution layers and 2 fully connected layers. Each input image is downscaled at seven levels via subsampling operations before being processed by a tailored network. The pooled feature maps from all channels are concatenated before a pixel-wise convolution layer. Then the feature maps from this last convolution layer are passed into the fully connected classifier. Many deep learning-based pattern recognition systems for biological-image classification tasks are based on transfer learning or reusing a pre-trained model. This methodology is extremely effective both for reducing design time and performance improvement, especially for the limited dataset. The possible ways of application include: fine-tuning the DNN without major modification of its structure, combining DNN with hand-crafted features or ensembling the feature outputs generated by multiple DNNs. Each strategy has its advantages and disadvantages. Firstly, fine-tuning an available pre-trained CNN is the most feasible approach and could be quickly adapted to a wide range of applications. However, the obtained results strongly depend on the similarity between the target dataset and the dataset for pre-training the reused network. Secondly, integrating hand-crafted features could exploit specific features of the target dataset and thus increases the accuracy but with the burden of designing an invariant feature set which usually requires expert knowledge of the domain. Finally, ensembling could be useful in most cases but the computation complexity is heavily increased. For example, in [79], three different pre-trained deep CNNs are jointly employed to perform feature extraction and the extracted feature vectors are concatenated before being used to train the two fully connected layers for the classification task. The size of the full network and the required number of calculations are at least three times bigger than the individual one. Unsupervised feature learning is also applicable for microscope image classification but with limited reports in the literature. The strategy is based on an autoencoder architecture, to learn visual features from the raw microscope image. In [80], based on vanilla autoencoder, a sparsity constraint is additionally introduced to the model during the training process to reconstruct the original image. This help to better capture specific visual features like edges besides general color and texture patterns as can be obtained in other feature extraction method (in particular, Discrete Cosine Transform). The sparse weights are then used as convolution filters on the input image to build feature maps being condensed by an average pooling. Finally, a softmax classifier is trained in a supervised manner to specify if the input image is cancerous. Another improved variant of sparse coding to efficiently compute large-scale features is applied in the classification of distinct components of tumor histology sections [81].

2.3 Generalization in deep learning

The performance of any learning model is evaluated by the ability to generalize from limited training examples to new data which was not used to train the model [82, 83]. This generalization concept resembles the learning process of humans, in which knowledge is transferred from one problem to the other [84]. For example, pathologists learn the properties of cells in representative tissue samples and apply the knowledge to give diagnosis to other samples. Supervised deep learning is currently still the most popular approach and has achieved significant success in many applications of computer vision. In this method, the test error on unseen data is a measure of how well the learned model generalizes [85, 86]. Figure 2.10 shows a general schematic of training a deep learning model and corresponding errors. As shown in the figure, supervised learning algorithms rely on the assumption of independent and identical distribution (I.I.D) between testing and training data, which requires the future data to be distributed identically like the data used for training the model. Note that the true distributions for training and testing data are hidden and we can only obtain the data samples. Thus, after a deep learning model is effectively trained in one training domain, the I.I.D assumption may not hold and its validity in the testing domain is lost.

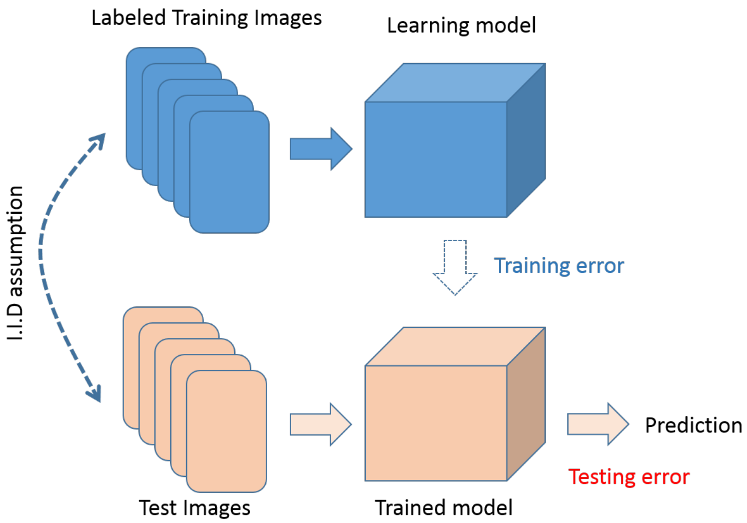


Figure 2.10 General schematic of training a deep learning model.

To measure the generalization performance of an algorithm, a simple way is to measure its performance on a held-out testing set containing unseen examples. An algorithm that works well on the training set but fails to generalize on the testing set is considered to be overfitting. Figure 2.11 shows the typical training error and test error as a function of the amount of training data [87]. The generalization gap is the difference between the two types of error and reflects how much worse the performance would be on a new data set compared to the training data. The goal of training deep learning algorithms is to reduce the test error or generalization gap. As the number of available training data points is small, the sampling noise is significant and the training data set becomes less representative of the true distribution that the future data is drawn. On the other hand, in a large training set, it is more likely there will be a related training sample for any particular test example while less likely there will be false regularities consistently appearing across samples [88]. In addition, it is harder for the network to memorize a large training set than a small dataset. Therefore, the restricted data available at training time limits the generalization of deep learning methods.

Another aspect to reason qualitatively about generalization is the model capacity, which roughly corresponds to the number of trainable parameters [88]. A network with little capacity is more likely to be underfitting, or less likely to fit all the regularities of the data in

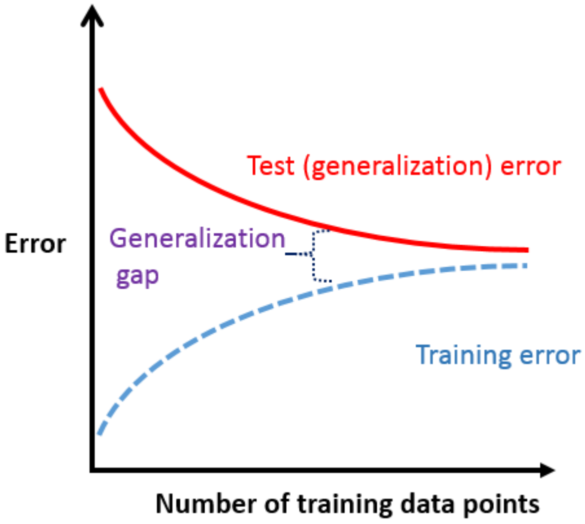


Figure 2.11 Relationship between generalization error and the training set size.

the feature space. However, if it has too much capacity, it can store or memorize the correct prediction for every training example and fail to generalize on unseen testing examples. As shown in Figure 2.12, it is necessary to design network architectures with a balanced capacity to have good generalization.

2.3.1 Techniques for improving generalization

This section describes existing techniques for improving deep neural network generalization. Most deep learning methods combine several of them in practice. It is worth highlighting that not every technique has theoretical justification and the use of any single or multiple techniques does not guarantee good generalization [88].

Data augmentation

This is one of the most common techniques used in deep learning methods due to its simplicity and high effectiveness. Given a limited training set, we can artificially increase its effective size by applying a set of random distortions or transformations to the inputs to the deep neural network. The transformations are usually considered as a way of simulating the domain shift and thus improve the generalization [89]. They could be simple conventional image transformations or be modeled with deep learning-based networks. The conventional transformations could be used alone or in combination, such as translating, rotating, zooming, color jittering, adding noise or warping, but their effects depend on the types of application

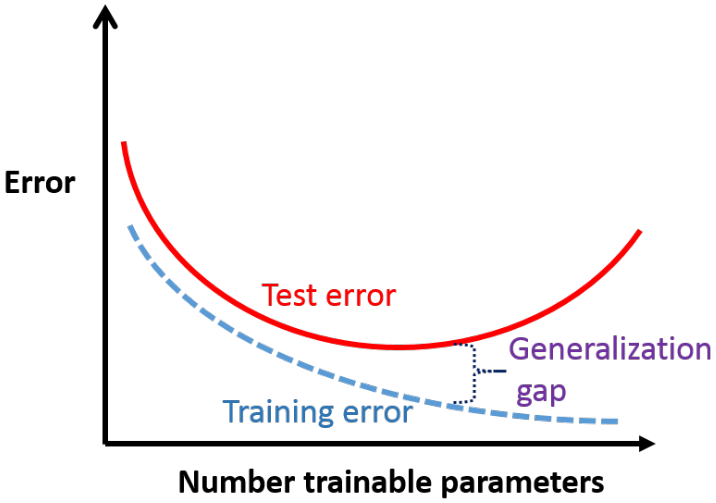


Figure 2.12 Relationship between generalization error and the number of trainable parameters.

and images. For example, randomly distorting the colors of objects inside stained brightfield images might decrease the performance as the objects in augmented image have inverted colors with regards to their true stained colors. The data augmentation could be done in an offline manner (augmented data are generated and stored before training the networks) or online manner (random transformations are applied on every input at training time). On the other hand, model-based augmentation usually relies on an image-to-image translation model, for example, CycleGAN [90], which transforms images from one domain so that they have the style or characteristics of images from another domain. So the diversity of training images is increased as each image sample can have appearances in different environments.

Regarding the lack of training data, various transfer learning techniques are also considered, such as training the DL model on large available image datasets before retraining it on the target dataset(“transferring the knowledge”) [91].

Reducing capacity

As suggested by the typical relationship between generalization error and model capacity in Figure 2.12, a direct strategy could be designing architecture without too high complexity. This could be aided by using the relevant number of layers, the number of neural units per layer [88], using compact structures like depth-wise convolution instead of standard convolution [92] or the bottleneck layer [93]. To find a balanced architecture that has enough capacity to learn distinguishing features but can avoid overfitting, we can rely on the validation set to tune the hyper-parameters. However, in practice, if we start by designing a simple network, this network could be too simple with regard to potential approaches. Some following techniques tend to maintain high capacity and prevent overfitting.

Regularization

Regularization is the technique that adds one or several regularization terms or regularizers into the training cost function. The total cost function is then the sum of the average loss and regularizers. Intuitively, regularizers penalize hypotheses that we think are unlikely to generalize well or unstable. For example, a cost function with R_{L_2} regularization to train a linear regression model is represented by:

$$C(\theta) = \frac{1}{N}\sum_{i=1}^{N} L(y(x,\theta),t) + R_{L_2}(\theta)$$

where

$$R_{L_2}(\theta) = \frac{\lambda}{2}\sum_{j=1}^{D} w_j^2$$

In this case, R_{L_2} penalizes the sum of squares of all the weights of the network and favors hypotheses where the norms of the weights are smaller. If two hypotheses can fit the training set, the hypothesis with smaller weights will produce a small loss while the other is producing a bigger loss and is probably more sensible when data distribution changes [88].

There exists another type of regularization which injects some stochasticity into the network computations to prevent overfitting, for instance, the dropout technique [94]. Instead of adding a regularization term to the cost function, it turns off each neural unit with a probability ρ. This is equivalent to multiplying the activations $\phi(z^{(i)})$ in a layer l_i with a randomly distributed binary mask m_i:

$$l_i = m_i.\phi(z^{(i)})$$

Note that at test time, all units will be turned on to avoid stochastic prediction and thus the weights are scaled by $1-\rho$ to compensate for the latent signal level. The dropout technique is more frequently applied to fully connected layers because they contain significantly more trainable parameters than convolution layers and thus make the network more susceptible to overfitting. Some other stochastic regularization effects could also be observed in the optimization methods, such as batch normalization [95] or stochastic gradient descent [88].

Early stopping

This technique is applied during the training process and can be represented by Figure 2.13. In the training, a validation error is considered to represent a testing error or generalization error. Thus, this technique aims at finding the training point where the validation error starts to increase. In practice, due to the fluctuation of training and validation error during training, it is not reliable to obtain this starting point. There is usually a heuristic alternative, in which we define a " threshold τ" and a "patient period δ", and we will stop training if after this period δ, the validation error can improve less than τ.

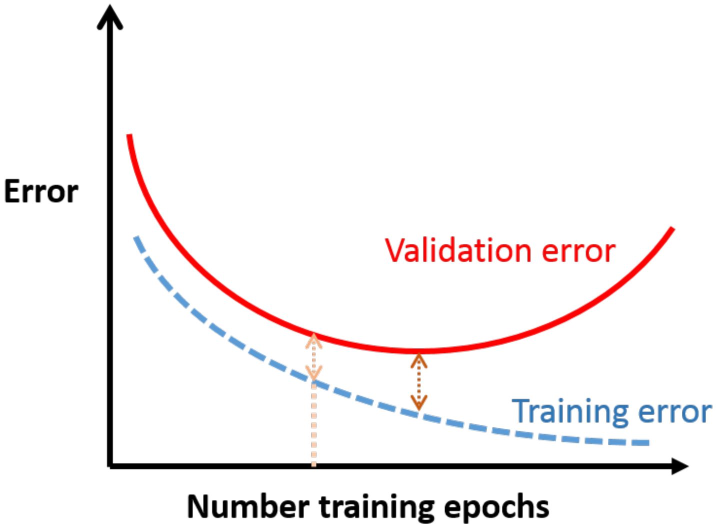


Figure 2.13 Relationship between generalization error and the training epochs.

Ensemble learning

This technique is based on the observation that the average prediction of multiple networks trained independently on separate training sets has lower variance than individual networks and an ensemble of networks often generalizes a bit better than single networks [88]. In practice, there are some effective approaches to producing an ensemble of networks, such as:

- Train the same deep neural network on random subsets of the full training data set.
- Train different deep neural networks on the same training data set.
- Combine deep neural networks with other learning algorithms or transformations.

This technique has been shown very effective in boosting the performance of DL model in various computer vision applications.

2.3.2 Application of generalization techniques in microscopy image analysis

In this section, we review existing techniques that were developed in DL methods for microscopy image analysis.

Data manipulation

DL methods are categorized into three main groups: unsupervised, partially supervised and supervised learning with many of them currently using supervised learning as they usually bring the highest performances [91]. However, the deep learning models have already experienced difficulties in achieving good generalization when training with a limited labeled dataset of microscopy images. A direct method is to create a large and diverse public database to be reusable in different applications. For example, the Cancer Genome Atlas (TCGA) database provides pixel-wise annotations for nuclei from different organs [96]. Some other specific datasets are also undisclosed to the public in competitions [97]. However, this approach requires an enormous effort and is not always readily available in many practical cases. Thus, in most of the published articles, data manipulation has been studied to deliver a sizeable set of representative samples for training DL models, usually with millions of parameters. We can categorize existing works into two major types: data augmentation based on conventional image transformations and data generation based on DL models. Authors in [98] performed comprehensive experiments on the sampling and data augmentation methods together with their related parameters and analyzed their impact on the generalization accuracy of microscopy image segmentation. They concluded that rotation, reflection, and jitter are the best transformations to reduce the generalization error gap and are hypothesized to closely mimic the visual variations. In practice, data augmentation is applied for training most DL segmentation methods across a wide range of applications, including segmentation of cells [11, 99], organelles [100, 101] or vessels [102]. In most deep learning-based classifiers, the augmentation using image flipping, rotating, or intensity altering is a typical strategy to achieve an accuracy of more than 90% [12], such as in cancer detection [103] or cell type classification [68, 73, 104].

Other works propose to create a supported training dataset by data generation strategies. For example, to segment nuclei in bright-field images, the authors in [105] design a workflow to obtain both the bright-field and the corresponding fluorescent images of the samples which enables generating ground-truth easily for training the DL network. To alleviate such laborious procedure, many works rely on some generative models such as Variational Auto-encoder (VAE) or Generative Adversarial Network (GAN) to generate simulated data. For example, CycleGAN is a popular style transfer model which is used to convert images from one source domain to the target domain. With the source dataset that is easy to obtain the ground truth (for e.g fluorescent images), a new dataset of images similar to target images could be synthesized and used as annotated training data [106, 107]. The main drawbacks of this method are that the adversarial network is difficult to train and usually has high computational complexity. In addition, the synthetic images may have low quality and many artifacts. Thus, further processing steps or techniques need to be applied, such as domain adaptation methods. Recently, the self-supervised learning method has also been studied for nuclei segmentation in histopathological images [108]. The self-supervised learning involves generating free labels from the original data to pre-train the DL network, referred to as a pretext task. It is supposed to improve generalization because solving the pretext task enables a network to learn generic features independent of the target task and reduce the overfitting to domain-specific regularities [89].

Domain adaptation

Due to many factors such as illumination, staining or image quality, there is a difference in distribution between the images used for training the DL network (source domain) and the images that need the prediction from the trained network (target domain). This domain shift probably degrades the performance of any DL model as they are designed with the assumption of Identical Independent Distribution (IID). Domain adaptation (DA) methods focus on reducing the discrepancy between the source and target domains by leveraging unlabeled target data and could be viewed as an extension of transfer learning.

Without the need for target data labeling, it would be beneficial for pathologists in real-world scenarios. However, this method has been barely studied for microscopy image analysis [107]. [109] is the first work that uses domain adaptation for epithelium-stroma classification in histopathological images. Their basic idea is to search for a reconstruction coefficient matrix to transform the source and target data into a common space where data in one domain could be reconstructed linearly by using the data in the other domain and then adjust the convolutional kernels for processing the images in target domains. In [110], for the classification of prostate histopathology images, the adaptation relies on adversarial training to minimize the distribution discrepancy in feature space between the two domains. In this architecture, the two feature spaces are generated from the source and target network, where the source network has already been trained using supervised learning. By jointly training the discriminator and target network using the GAN loss, the target network could extract domain invariant features. The methods proposed in [111] and [107] are designed for domain adaptation in nuclei segmentation. Both of the works are based on GAN architecture and use synthetic histopathology images for data augmentation. While the first work only considers pixel-level adaptation, the latter minimizes also the domain gap in the feature levels.

There are two major disadvantages of DA methods [89]: (1) the model adaptation or fine-tuning has to be repeated whenever there is a domain change, and (2) the assumption that target data are readily available may not be realistic in clinical settings as the image samples of a future patient are not known before the deployment of the DL model.

Network design

In the domain of microscopy image analysis, designing a novel model architecture is challenging, and requires computational knowledge which is uncommon among biomedical scientists. Most existing DL methods focus on training strategies, such as data manipulation and transfer learning, to make the model more generalizable. In practice, some computer vision-winning architectures such as ResNet [112] or Mask R-CNN [113] have been commonly reused as the backbone in many works. The common perception is that these models should also be able to produce high performances on almost any type of digital image. However, there exists a few works that focus on designing novel network architectures for improving the learning generalization [89]. They demonstrate that specially designed architecture with an adapted optimization strategy could work better than pre-trained winning models. For example, the model MicroNet, recently proposed in [114], extends U-Net [6] by designing a supervised multi-resolution network architecture. The network processes the input at multiple resolutions, maintains intermediate connections between layers, and generates the output using multi-resolution deconvolution filters. The weighted loss function is calculated by combining the losses of all the branches. The batch normalization technique is applied to all the branches during training. The proposed network could be used to segment different objects in fluorescence microscopy and histology images such as cells, nuclei and glands after specifically modifying the network hyper-parameters. However, it relies on stain normalization to reduce the effect of staining variation across laboratories and conditions. In addition, several special data augmentation techniques were used, that require a visual examination to ensure the distortions are realistic and not overshoot. Authors in [115] also proposed a multi-scale network, SAMS-NET, that is robust to stain variations in $H\&E$ images. In this work, they introduced a loss function that includes a pre-defined weighted map that is sensitive to the Haematoxylin intensity and important areas in the image. In other methods, authors designed architectures that can exploit more supervised information from the labeled dataset. For example, DCAN [116] is a dual architecture that outputs the nuclear cluster and the nuclear contour as two separate prediction maps. [117] proposed a network with a custom weighted cost function based on the relative position of pixels within the image to improve and stabilize the prediction of the inner nuclei and contour. Recently, authors in [118] designed a multi-head network to exploit the vertical and horizontal distances of nuclear pixels to their centers of mass.

Ensemble learning

Ensemble learning was also studied extensively to improve the performance of DL classifiers or segmentation models in microscopy image analysis. In these designs, the component networks could be combined in different ways. For example, to take advantage of transfer learning, the ensemble model could be composed of multiple pre-trained networks, each of which processes the same input images or input images of different scales. This simple technique has been shown very effective to boost performance across a wide range of applications. In [79], the authors designed a model that concatenates the features extracted from three different pre-trained networks before feeding them to the fully connected layers to perform the classification of sub-cellular organelles fluorescent images and pap-smear bright-field images. To identify phenotypes in cellular images, the M-CNN architecture developed in [78] processes multiple scales of an input image over seven parallel convolutional pathways and then concatenates the output feature maps to feed into a final convolutional layer. In a recent method aiming at breast tumor nuclei segmentation, described in [119], the authors designed a two-stage network to first concatenate feature maps from three U-Net like DCNNs (each has either VGG-19 [120], DenseNet-121 [121], or ResNet-101 [122] encoder), then use the concatenated images as input to an additional U-Net [6].

CHAPTER 3 RESEARCH OBJECTIVES

3.1 Problem Statement

Recently, deep learning or deep convolution neural network is emerging as a powerful method that can be applied to various tasks of biological-image analysis, including segmentation and target classification [29, 123]. Compared with conventional machine-learning approaches, deep learning could directly process raw image data to automatically learn optimal features of objects' representation in an image. Thus, it helps to avoid the burden of hand-crafted feature engineering that requires much domain expertise and is inherently biased by designers. Even though deep learning models could produce significant improvement in certain applications, it still needs great effort to address the unique challenges for microscopy image analysis. To achieve top performance, many deep learning methods are based on either increasing the depth of neural network [38], integrated complex modules in parallel at each layer [124] or even combining multiple deep networks as in multiscale architectures. This strategy increases the number of learnable parameters or degrees of freedom. Thus, it increases model complexity and the risk of over-fitting the training dataset. As microscopy image datasets usually have limited annotated data, the deep learning-based methods are more susceptible to over-fitting problem and lack the generalization ability to achieve satisfactory results on unseen images. On the other hand, designing an efficient algorithm to generate additional annotated data or assign the correct label to a non-annotated image for enlarging training data to improve generalization capability is still a challenging problem. In many cases, the limited data possibly make the deep learning model have lower performance than conventional hand-crafted machine learning methods. For example, in the case of HeLa cell organelles classification, the state-of-the-art top-performance deep neural networks can not outperform conventional hand-crafted based feature extraction methods [125]. As a result, the application of deep learning models as common algorithms in supportive diagnostic tools in a practical scenario is strongly impacted.

There are two principal challenges preventing the deep learning algorithms from reaching a good generalization ability in microscopy image analysis: the scarcity of labeled data and high data variability [126]. Firstly, there is usually a lack of labeled data because manual annotation is a time-consuming and costly process that requires the expertise of the field. Secondly, microscopy images obtained from different experimental conditions present significant appearance variation because each sample preparation and scanning procedure has a wide range of parameters within or across laboratory settings together with underlying biological variability [15]. Usually, a finite number of training images are acquired under specific conditions that are likely different from those of target test images. This means that the testing domain in feature space differs from the training domain, or there is a "domain shift" (Figure 3.1).

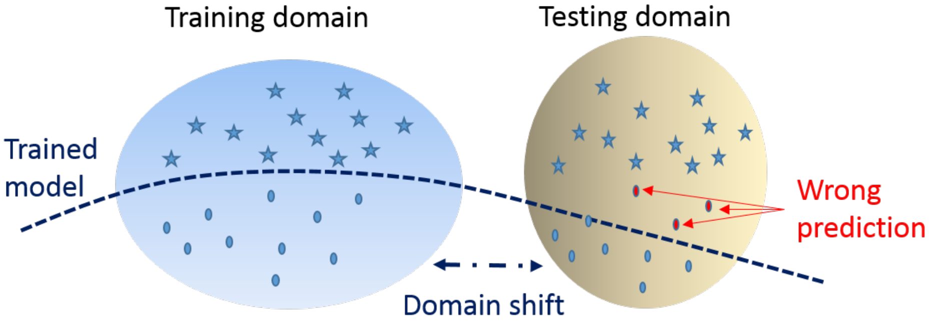


Figure 3.1 Illustration of data domain shift due to high variation [9].

Improving the generalization for microscopy image analysis in practical conditions is a challenging problem. As any finite set of training images only represents specific aspects of images that are possibly observable, any standard supervised deep learning is likely to not generalize well on unseen images. An ideal solution should be able to extract domain invariant features from finite labeled training samples in the source domain that can work well in the target domain. A trivial strategy that exploits additional source domains or data augmentation may not enhance or even degrade the generalization of a model [126]. There exist two groups of machine learning methods to directly deal with the appearance variability of microscopy images: pre-processing image data and modifying the learning model with a relevant training mechanism [15]. The first category focuses on staining normalization, data augmentation through color transformations or a combination of the two techniques. Methods of the second category are mostly based on domain adaptation in which a learning model learned firstly from a source domain will be fine-tuned lately to adapt to a new target domain. The adaptation will be done by retraining entirely or partly the model on samples of target data. So, this method requires collecting data from target domains and going through the process of retraining the model. While the manual pre-processing image techniques are specific to each application and produce limited performance gain, the domain adaptation approaches require collecting future target samples, which are usually unavailable in clinical settings, to adjust the trained model before the deployment.

Thus, more efficient deep learning algorithms, that can automatically extract the features from raw images and exploit well available data to improve generalization, are still much needed in practice.

3.2 Specific Objectives

The overall goal of this research project is to improve the generalization and applicability of deep learning in the segmentation and classification of microscopy images. To achieve our goal, we establish the following objectives:

3.2.1 Objective 1

We develop an unsupervised segmentation model that is domain-adaptable. In particular, we focus on the automatic segmentation of nuclei from the cytoplasm, where samples are stained with hematoxylin and eosin agents, as this is a common task in many histopathology procedures. Existing models typically offer high performance but are heavily dependent on expensive manual annotation. The benefits of domain-adaptability make the design suitable for deployment in a variety of histopathology settings.

3.2.2 Objective 2

We develop new classification models to deal with the problem of limited labeled microscopy images acquired from different imaging domains. The algorithms are expected to have good generalization given small training sets, requiring low complexity and achieving high performance on unseen test images. This problem is regarded as weakly supervised training and is currently still a challenging problem for any deep learning model.

3.3 General methodology

3.3.1 Segmentation

In order to develop unsupervised learning segmentation model in objective 3.2.1, we design a solution based on two following strategies. Firstly, physical aspects of the sample preparation and image acquisition process were investigated to extract prior knowledge which are useful for the model design. Knowing the visual appearance of nuclei and cytoplasm after staining helps to develop a pipeline to generate synthetic data for training the deep learning network. Moreover, the optical density parameters enabled us to perform the color separation between them. Secondly, we attempt to build modular design where parallel modules process different subspaces of features and provide a user-adjustable controlling parameter to combine the output feature maps. We use a data-centric approach where the available information from target data is extracted during the network training, which benefits the domain adaptability. Unlike other domain-adaptable methods, we do not collect a set of realistic images in the target domain to adapt the DL model before its deployment.

To evaluate the performance of the developed model and compare with existed works, we use different datasets that are publicly available. The images from these datasets are collected from multiple hospitals across several countries, ensuring the diverse appearances for testing purpose. The performance metrics are recorded following the common formulas in recent works. This first contribution was submitted to the Journal of "IEEE Transactions on Medical Imaging" and is presented in chapter 4.

3.3.2 Classification

For objective 3.2.2, we first explore the combination of compact structures to design a model for learning microscopy images acquired from different imaging devices and with various objects of interest. This is the first work that elaborate a light weight model for multi-domain learning of microscopy images and without requiring the adaptation of domain-specific parameters. Microscopy datasets are usually small and have sparse non-identical distributions. As the number of labeled samples per class is small, the distances between feature spaces in different domains are large, especially when both the visual appearance and object of interest differ. Therefore, it is difficult for the deep learning network to learn unified representation, or converge to an efficient feature set, across various datasets.

In addition, we will develop new regularization techniques for the training of supervised convolution neural networks. A standard approach to overcome the lack of labeled data is data augmentation. However, it is sensitive to select the types and ranges of transformations to be suitable for each application and it is unclear how to control the outlier caused by mistaken augmented data. The regularization criteria will be integrated into the learning objective such that the algorithm could be trained in an end-to-end fashion. This second contribution was submitted to the journal "Computer Methods and Programs in Biomedicine Update" and is described in chapter 5.

Then, instead of the above generic model, we focus on developing deep learning architectures that can outperform state of the art models in a specific application, particularly the fluorescent microscopy images. These models are expected to produce higher performance than the previous generic model. As the first solution to achieve this goal, we design a model combining lightweight convolution neural network structures with multi-resolution analysis. In each decomposed spaces of an input image, convolution kernels will extract discriminative features and provide various pattern characteristics of the same organelle. This third contribution was published in the conference proceedings of “International Conference on Image Analysis and Recognition ICIAR 2019” [125] in August 2019 and the detail is provided in chapter 6.

In our second approach, we develop an efficient model based on a compact convolution neural network and deep embedded clustering. The operation of the model differs from other works in that prediction is performed by unsupervised clustering in feature space instead of producing class probability by the softmax function. We also formularize a regularization method to support the clustering technique. This fourth contribution was published in the conference proceedings of “42nd International Conference of The IEEE Engineering in Medicine and Biology Society EMBC 2020” [127] and is presented in chapter 7. The adopt of regularization technique for optimization was inspired by our another work published in the conference proceedings of “Pattern Recognition - ICPR 2020 Workshop of Deep Learning for Pattern Recognition” [128] and is described in Annex A.

Unlike common methods for microscopy image analysis, which exploited existing pretrained heavy DL networks and extensive data augmentation to get better performance, we elaborate new approaches to surpass their performances even without data augmentation or with extremely less available training data.

For performance evaluation and comparison, we obtain data from multiple public datasets. Especially, to assess the multi-domain classifiers, we use datasets generated by different laboratories and the images represents samples at different cellular levels, including tissue, cells and organelles. We quantify the performances by recording the popular classification metrics.

CHAPTER 4 ARTICLE 1: INTERPRETABLE MODEL FOR NUCLEI SEGMENTATION IN HISTOPATHOLOGICAL IMAGES USING UNSUPERVISED LEARNING

Duc Hoa Tran[1], Michel Meunier[2], Farida Cheriet[1]

[1]Department of Computer and Software Engineering, Polytechnique Montréal, Canada
[2]Department of Engineering Physics, Polytechnique Montréal, Canada

Presentation This work was submitted to the Journal of "IEEE Transactions on Medical Imaging" on March 6, 2022.

4.1 Abstract

Existing deep learning (DL) models for the segmentation of nuclei in histopathology images typically offer high performance but are heavily dependent on expensive manual annotation. Even when annotated public datasets are available, their generalization capability and operation explainability are noticeably limited. In this work, we present an unsupervised DL model for nuclei segmentation in histopathological images that is interpretable and domain-adaptable. Our approach is composed of two strategies: (1) integrating relevant prior knowledge of image acquisition conditions into the model's design, and (2) flexibly combining a classical algorithm with a DL-based encoder-decoder network. We also focus on exploiting available information from each target image during training of the DL network. This is achieved by generating a synthesized training dataset that is visually similar to the target image, and by using a generated approximate mask for the target image to regularize the network parameters. We evaluated the performance of our method on three public datasets of histopathological images used in recent challenges for nuclei segmentation. The results demonstrate that our method can outperform other unsupervised segmentation approaches and produce results that are comparable with supervised DL models.

4.2 Introduction

Histopathology images provide helpful data that pathologists need to analyze before grading the stage of various diseases, in particular cancer. There will be many difficulties if this assessment is done manually, considering the high complexity and variation of sample images.

Computational histopathology alleviates these limitations by adopting digital image processing and computer vision techniques to analyze the images automatically. This approach has been demonstrated to improve the performance and throughput of many analysis tasks, including detection, segmentation, and classification of different objects of interest [129]. In this study, we focus on the automatic segmentation of nuclei from the cytoplasm, where samples are stained with the widely used hematoxylin and eosin ($H\&E$) agents, as this is an essential step in most histopathology procedures. Pathologists can obtain various morphological and appearance metrics, such as size, pleomorphism, distribution, and nucleus/cytoplasm ratio, that are important for disease diagnosis and prognosis [130].

Nuclei segmentation in $H\&E$ histopathology images is challenging for several reasons. The first challenge is the sparse and unclear color separation between nuclei and cytoplasm regions. Eosin is an acid that turns the basic constituents of the cytoplasm and collagen fibrils pinkish. In contrast, Hematoxylin dyes the acidic components of the cell a bluish color. As the most acidic components are deoxyribonucleic acid (DNA) and ribonucleic acid (RNA), not only the nucleus but also the RNA-rich regions of the cytoplasm and the surrounding matrix of cartilage produce dark blue or purple color [131]. The wide variation in cell morphology is another difficulty for the segmentation task. Not only do cell nuclei in different organs tend to have different sizes and textures, but nuclei of the same cell type can also vary in appearance.

Image processing techniques for nuclei segmentation are based on a few popular algorithms, such as thresholding, watershed, active contours and graph cuts, to exploit the distinct color and morphology of the nucleus within a cell [108, 132]. These classical algorithms are theoretically interpretable because they comprise a fixed set of rules with certain preconditions to ensure the desired outputs. However, they cannot generalize well in real-world situations since histopathology image data often do not conform to the preconditions, considering diverse tissue types and appearances.

On the other hand, data-driven methods exploit characteristic features extracted from real images to produce higher performance. Early machine learning approaches rely on manually designing a diverse set of nuclei features such as color, texture variance, shape, and intensity gradients [133]. This feature engineering process involves mapping from visual features represented in a dataset to theoretical formulas. Thus they are time-consuming and highly specific to a given target dataset. Modern Deep Learning (DL) approaches are more powerful in practical scenarios as they are designed to learn distinguishing features directly from raw image data. Unfortunately, DL algorithms work under the assumption that the training and testing data are independent and identically distributed. In practice, they cannot generalize well beyond the data examples that were used to train the network. Moreover, most existing approaches use a supervised strategy, which requires manual pixel-wise annotations by experienced pathologists. The annotation process to create a ground truth image dataset is time-consuming and often marred by high variability between observers.

Motivated by the lack of segmentation methods that don't depend on annotated training sets, we propose an interpretable and adaptable model based on an unsupervised approach to segment the nuclei in histopathology images. In our approach, rather than use a black-box DL model (and attempt to explain its performance), we choose instead to design an interpretable solution, by combining the following strategies:

- We leverage physical parameters in the imaging process that are known or can be obtained experimentally, such as the transformation matrix for color deconvolution of staining agents.
- We propose a modular design where two parallel modules process different subspaces of features; output feature maps are then aggregated by a user-adjustable controlling parameter. As users can evaluate which subspace of features is more important than the other, this enhances model interpretability.

To address the challenge of generalization in deep learning-based models, we use a data-centric approach where the available information from target data is extracted during the network training. In particular, we present a simple but effective method to estimate the background textures and the stained nuclei objects directly from the target image, and to use this information to synthesize the training dataset. In contrast to other domain-adaptable methods, we do not need to collect a set of realistic images in the target domain to optimize the DL network before its deployment.

In summary, our main contributions in this study are:

- Designing an unsupervised deep learning framework for nuclei segmentation without training data annotation.
- Proposing a novel strategy for interpretable model design by using physical parameters and user-controllable hyper-parameters.
- Developing an integrated optimization function to improve network adaptability to the target image.

This paper is organized as follows: we summarize related works in section 4.3, then describe our methodology in section 4.4. In section 4.5, we present the three public datasets and evaluation metrics used in this study. We report the different experimental results in section 4.6. Finally, we discuss the impact and conclusion of this work in sections 4.7 and 4.8.

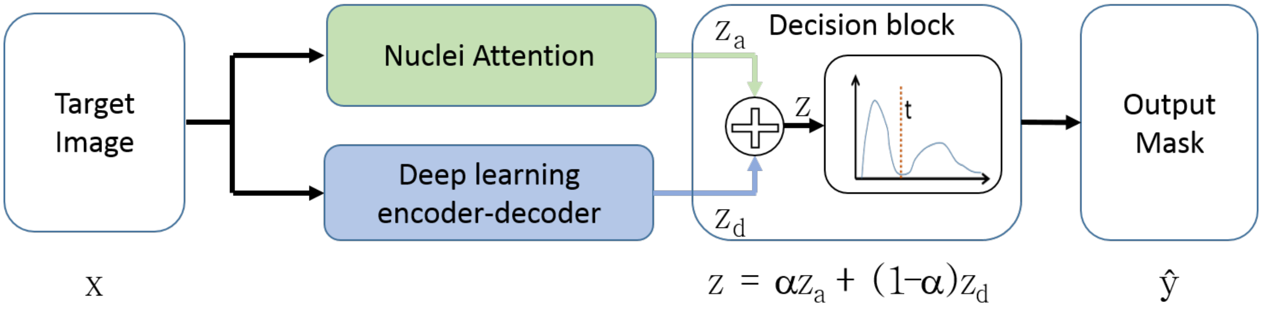


Figure 4.1 Inference mechanism to produce nuclei segmentation mask.

4.3 Related Work

Most published articles on automated nuclei segmentation use single image processing algorithms, or a combination thereof, among intensity thresholding, morphological filtering, region growing, watershed, clustering, deformable model fitting, and graph cuts [130] [134] [135] [136]. These methods lack robustness to the appearance of nuclei in histopathological images, such as inter-nuclei and intra-nucleus color variations or color diffusion between nuclei and background regions.

Deep neural networks, i.e. deep learning, has become the dominant trend for research in microscope image analysis since 2015 [22]. Unlike methods based on hand-crafted feature extraction, DL algorithms can extract optimal discriminant feature representations directly from raw image data by optimizing a cost function. A comprehensive review in [137] reveals that Convolutional Neural Networks (CNNs) and Fully Convolution Neural Networks (FCNs) are the most popular architectures for microscopy image segmentation. In some cases, other architectures such as recurrent neural networks and stacked autoencoders have also been used.

Most state-of-the-art nuclei segmentation methods apply supervised training when annotated images are available to guide the optimization of the DL model [108]. However, due to the limited size of these datasets, with just dozens to hundreds of images, training a DL model generally suffer from serious overfitting or low generalizability. In this situation, a practical solution reported in many publications is transfer learning. Specifically, the feature extraction module of a network previously trained on a large natural photographic dataset, such as ImageNet [138], is fine-tuned to extract useful features from objects of interest in a microscopy dataset with limited annotations.

Several strategies to deal with the training data requirement have been proposed. A straightforward method is to create a large and diverse public database to help produce DL models that are reusable in different applications. For example, the Cancer Genome Atlas (TCGA) represents a substantial effort to provide pixel-wise annotations for nuclei from different organs [108]. The major disadvantage is that, before using the pre-trained models, the pathologist needs to verify that their own data have identical representations to those of the database used for training, which is very challenging to do [139]. This is because the representation similarity depends on various factors including microscope type, optical settings, and experimental preparation. Otherwise, the model will produce unexpected results that are difficult to assess visually.

In unsupervised domain adaptation (DA) methods, it is assumed that a certain number of unlabeled images in the target domain can be collected and that there exists a large set of annotated images referred to as source domain images [107]. A domain refers to a set of nuclei images that are visually similar and are captured in the same conditions. During training, a DL network is first optimized with labeled images in the source domain, then it is adjusted via an estimated relationship between unlabeled target images and source images, such as the distance between the distributions of extracted feature maps. However, in many cases, it is impossible to collect target data before designing and optimizing the DL network [89, 140]. In clinical settings, the images and their acquisition parameters for a new patient are not known in advance. Conversely, it is impractical to require future data to conform to predefined sample preparation and image capturing conditions. Other works have attempted to use generative DL models to directly translate images from a source domain to a target domain before training the deep network [141]. Although demonstrating high potential, these methods remain prone to uncontrollable artifact generation due to the high complexity of the training process [139]. Another promising method is the self-supervised approach proposed recently in [108]. To train a fully convolutional network that outputs a segmentation map, the authors cascaded it with another CNN which learns to classify the predefined magnification levels of input images. Although the classifier component provides a self-supervision signal for training without image annotation, an annotated validation set is still needed to determine the optimal model.

In summary, the existing methods cannot eliminate the dependence on labor-intensive and time-consuming histopathology annotation. They also provide no guarantee of generalization when deployed in the clinical setting because no information from the real target domain can be exploited when training the DL model. In this study, we propose a solution to these two shortcomings by developing an unsupervised and domain-adaptable model.

4.4 Method

In this section, we describe the core principle of our nuclei segmentation framework. We first summarize the general process of producing segmentation masks in section 4.4.1, then explain in detail the two main structures in sections 4.4.2 and 4.4.3.

4.4.1 Inference mechanism

As represented in Figure 4.1, we introduce two pathways for the inference pipeline, a nuclei attention module based on color deconvolution and a deep learning-based encoder-decoder structure. These pathways will produce two latent representations z_a and z_d, which are fed into a decision block. Here, they will be aggregated by a control parameter α (where $\alpha \in [0.0, 1.0]$) to produce a unified representation z.

$$z = \alpha z_a + (1 - \alpha) z_d \tag{4.1}$$

Our goal is to allow each pathway to produce a different perspective to contribute to the prediction map. Then, the automatic thresholding Otsu method [142] is applied to z to create the binary semantic segmentation. After that, we compute the distance to the background map and select the minima of the opposite of the distance as the markers for the watershed algorithm [143] to generate the final instance segmentation mask.

The controlling coefficient α has a default value, but the user can change it at inference time to adjust the performance. This design helps to address two limitations: (1) α is not a hyper-parameter to be defined before training the DL-based encoder-decoder model; (2) the α value most adapted to the target data can be explored after deployment.

4.4.2 Nuclei attention

As the first pathway, we generate a feature map that highlights the probable nuclei locations in the target image based on their true color. To do so, we perform color deconvolution on the target RGB image using the method proposed in [144] to separate the Hematoxylin channel, which provides better nucleus contrast. Considering a particular pixel location (i), there are different amounts of staining dyes. The optical density (OD_i) level in each channel of the RGB image is linearly proportional to the amounts of absorbing dyes, as follows:

$$OD_i = -\log(y_i) = I_{HED}K \tag{4.2}$$

Here, y_i is the image pixel values in RGB format. The vector I_{HED} contains the amounts of dyes at a given pixel, and K is the normalized matrix representing the OD of pure stains (a stain with a single dye). The authors of [144] provided K for the three most common dyes (Hematoxylin (H), Eosin (E), and Diaminobenzidine (D)) as follows:

$$K = \begin{bmatrix} 0.65 & 0.70 & 0.29 \\ 0.07 & 0.99 & 0.11 \\ 0.27 & 0.57 & 0.78 \end{bmatrix} \tag{4.3}$$

Each row of K represents the independent contribution of a stain H, E or D in each of the R, G and B channels. These stain-specific values are determined by measuring relative absorption for red, green, and blue on slides stained with each of the three dyes.

We can thus obtain the orthogonal representation of the dyes forming the original RGB image:

$$I_{HED} = -\log(y_i)K^{-1} \tag{4.4}$$

Here we are interested in the Hematoxylin channel in I_{HED} as it highlights the cell nuclei and use it as the nuclei attention map z_a.

4.4.3 Deep learning encoder-decoder network

An encoder-decoder module is used as the second pathway, to produce a feature map to segment the target image. To train the network, we want to exploit as much as possible the available information from the target image by performing two main steps: (1) synthesize a set of similar images with precise masks and (2) regularize the network parameters by the target image with the approximate mask. In this work, we used U-Net [6] layers for the encoder-decoder network, but other segmentation models could be employed in our framework. The training process is summarized in Figure 4.2.

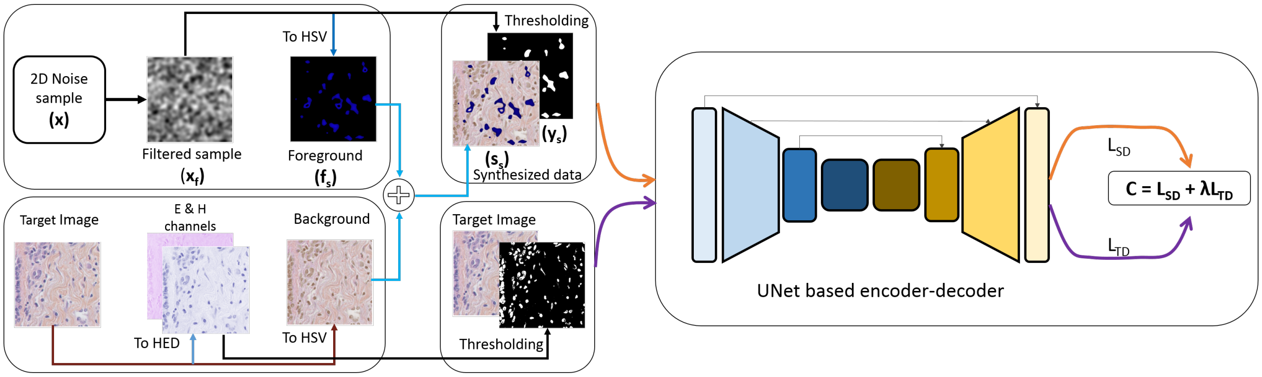


Figure 4.2 Training process for the deep learning-based encoder-decoder module.

Synthesizing target-domain dataset

In this first stage, we aim at creating a synthetic dataset from every single target image. Firstly, we randomly generate gray-scale images of nuclei objects and corresponding masks. Each sample pair is created by Algorithm 1. Our algorithm is inspired by the CowMask algorithm [145] which can generate random dropout regions for image augmentation when training deep learning networks for the classification task.

In brief, we first create a two-dimensional noise sample from a normal (Gaussian) distribution. Then we apply the smooth filtering in which the scale σ is drawn from a log-uniform distribution $\log\mathcal{U}$ in a pre-defined range $(\sigma_{min}, \sigma_{max})$. To define the maximum proportion of nuclei area in the sample image, we rely on a threshold τ which is defined via the inverse error function erf^{-1} of the normalized distribution. As demonstrated in [145], this threshold defines the intensity level below which the proportion of smooth noise pixels is p. The value of p is drawn from a uniform distribution $\mathcal{U}$ within the pre-defined range $(p_{min},\, p_{max})$ Therefore, we can limit the maximum proportion of nuclei pixels within the generated image u_s. To further control the scale of the nuclear blobs, we use another user-defined threshold T_m. After dropping the low-intensity pixels and normalizing the resulting image to the full dynamic range [0,255] as u_{sn}, a binary mask y_s is created by comparing each pixel of u_{sn} with T_m and used to extract the desired nuclei blobs x_s by element-wise multiplication with the normalized image. Note that the binary mask y_s of the nuclei is used later as the supervision signal for training our DL network.

Secondly, we transform the grayscale image x_s into a color image in Hue-Saturation-Value (HSV) space to simulate the color of Hematoxylin-stained nuclei. Compared to the Red-Green-Blue (RGB) color space, HSV isolates the Hue channel, which we exploit for color

Algorithm 1: Algorithm to generate each nuclei image and corresponding mask.

```
Input: size $H \times W$
Input: proportion range $(p_{min}, p_{max})$
Input: scale $(\sigma_{min}, \sigma_{max})$
Input: mask threshold $T_m$
Begin
    # Draw noise sample from normal distribution:
    $x \sim \mathcal{N}^{H \times W}(0, 1)$;
    # Smooth noise sample with a random sigma:
    $\sigma \sim \log\mathcal{U}(\sigma_{min}, \sigma_{max})$; $x_f = Filter(x, \sigma)$;
    # Compute mean and standard deviation:
    $m = mean(x_f)$; $s = std(x_f)$;
    # Set a random proportion:
    $p \sim \mathcal{U}(p_{min}, p_{max})$;
    # Inversely compute the threshold:
    $\tau = m + \sqrt{2}.\,\mathrm{erf}^{-1}(2p - 1).s$;
    # Drop pixels in proportion p:
    $u_s = x_f \geq \tau$;
    # Do binary thresholding to get the nuclei mask:
    $u_{sn} = normalize(u_s)$; $y_s = u_{sn} \geq T_m$;
    # Extract nuclei objects:
    $x_s = u_{sn}.y_s$;
End
Return nuclei blobs image $x_s$, nuclei mask $y_s$
```

processing. The converted nuclei image, referred to as foreground f_s, will have a random hue level within the blueish range while the saturation and value levels are left unchanged. Next, we transform the target image into a color image in HSV space and modify the entire hue channel to be pinkish while preserving its S and V channels. As a result, we obtain a background b_s that can be blended with the synthetic nuclei f_s. Our goal here is to exploit the background texture of the target image while camouflaging the real nuclei in the foreground, while training the encoder-decoder network. Finally, we combine the foreground and background using the following formula:

$$s_s = \eta * y_s.f_s + \bar{y}_s * b_s \tag{4.5}$$

where s_s is the resulting synthetic image. Each of the foreground (f_s) and background (b_s) images is respectively pixel-wise multiplied with the segmentation mask y_s and its inverse $\bar{y_s}$. The random factor η is added to adjust the intensity of the nuclei in the synthetic image. This acts as a data augmentation technique during training. We also randomly clear the interior of the nuclei to model chromatin phenomena before using this equation and apply a blurring filter to the synthetic image to soften the nuclei boundaries. Finally, the synthetic image is transformed from the HSV space back to RGB.

Regularizing the encoder-decoder network

Besides the synthesized image-mask pairs, we generate additional pseudo-supervised training data based on the target image. In particular, after producing the nuclei attention map as described in section 4.4.2, we segment the nuclei by Otsu thresholding [142]. The Otsu method may not precisely delineate nucleus boundaries and may include some non-nuclear material. This is acceptable because the objective of this step is to guarantee that the model will learn to minimize intra-class intensity variance, or equivalently, maximize the inter-class variance of image pixels [3].

This target image-mask pair is beneficial for optimizing the DL model in two ways: (1) by providing a supervision signal, thereby pushing the output performance to be better than conventional unsupervised algorithms; (2) by familiarizing the network with the representation of the target image, thereby reducing the domain gap between training and testing data.

The regularization is done by optimizing the weighted cost function in Equation 4.6, described in the following subsection.

Training procedure

Unlike most other methods, we train our DL network from scratch. In general, each training iteration comprises two steps: (1) training on a batch of synthesized image-mask pairs (see section 4.4.3) and (2) training on the target image with the approximate mask (section 4.4.3). Thus, the cost function is a weighted sum of the loss terms:

$$C = L_{SD} + \lambda L_{TD} \tag{4.6}$$

where L_{SD} denotes the average loss of each mini-batch of synthetic data in a training iteration, L_{TD} is the training loss on the target image with the generated mask, and λ is the weight to balance the two loss terms. In this work, we use the same loss function for both L_{SD} and L_{TD}. We followed the suggestion in [146] to define a hybrid loss function that combines soft Dice coefficient loss and pixel-wise binary cross-entropy loss, as follows:

$$L = 1 - \frac{2}{N} \sum_{1}^{N} \frac{\sum_{i=1}^{H\times W} p_i q_i}{\sum_{i=1}^{H\times W} p_i + \sum_{i=1}^{H\times W} q_i} + \frac{1}{2}\frac{1}{N} \sum_{1}^{N} \sum_{i=1}^{H\times W} q_i \log p_i \tag{4.7}$$

In this formula, N is the input batch size, and p_i is the predicted value for pixel $q_i \in \{0, 1\}$ in the target mask of size $H \times W$. Whereas the Dice loss handles the imbalance between nuclei and background areas within each image, the binary cross-entropy loss is an image-level indicator of segmentation performance and helps to maintain a smooth gradient change across training iterations.

We applied common data augmentation techniques on each input data during training, including random rotation, flipping, and contrast variation. On the other hand, we observed that random brightness and color jitter degraded the segmentation results. Note also that when generating the synthetic dataset as in section 4.4.3, we obtained randomly varied nuclei morphology and color values within the relevant range. This process is equivalent to offline data augmentation, but our augmentation is well controlled at the nuclei level instead of the whole image level as in standard augmentation techniques.

4.4.4 Implementation

We implemented our method in Python, using the PyTorch library [147] to build the U-Net segmentation network, and the Scikit-image [148] and OpenCV [28] libraries to synthesize histological images. To train the U-Net, we applied the SGD optimizer using Cosine Annealing schedule to set the learning rate [149], with initial learning rate $lr = 0.01$, weight decay

$\gamma = 0.001$, and momentum $\beta = 0.9$. After resizing the input images to 256×256, we used a batch size of 16 and applied data augmentation during each training iteration. We set the number of maximum training epochs to 10 whereas the network still relied on early stopping to determine the optimal model with the lowest validation error.

The overall model was trained in an end-to-end fashion, using a single NVIDIA V100 GPU. The entire processing time for each histological image, including the generation of the corresponding input data set and training the segmentation network, was less than 4 minutes.

4.5 Experiments

4.5.1 Datasets

Our proposed method was validated on three public histopathology datasets, namely MoNuSeg, CoNSeP, and TNBC, summarized in Table A.1. The Multi-Organ Nucleus Segmentation (MoNuSeg), in its original version, contains 30 annotated tissue images [130], each of which was extracted from a whole slide image (WSI) of an individual patient in the Cancer Genome Atlas (TCGA) database [96]. These images have size 1000×1000 at 40× magnification, capturing cell nuclei in 7 different organs (breast, bladder, colon, kidney, liver, prostate, and stomach) of patients in 18 hospitals in the USA. Subsequently, 14 images were added to MoNuSeg, comprising two more tissue types (lung and brain). It is considered to be the largest repository of manually annotated nuclei data [97].

The CoNSeP dataset consists of 41 stained images extracted from colorectal adenocarcinoma whole slide images of 16 patients in University Hospitals Coventry and Warwickshire, UK [118]. Unlike MoNuSeg, this dataset focuses on a single cancer type to better cover various visual fields of different tissue components and nuclei types.

The Triple-Negative Breast Cancer (TNBC) dataset is a collection of 50 annotated image patches, with 512×512 dimension at 40× magnification [150]. These images were collected from 11 different breast cancer patients at the Curie Institute in France.

In practice, it takes about 2 minutes to annotate each nucleus; it can thus take hundreds – even thousands – of working hours for a pathologist to generate these datasets [141]. Thus, unsupervised segmentation can offer great time and cost savings.

4.5.2 Parameters for nuclei image synthesis

In the nuclei generating algorithm for image synthesis, we set the scale parameter range $(\sigma_{min}, \sigma_{max}) = (5.0, 20.0)$ and mask threshold $T_m = 127$ for all datasets. As there were

Table 4.1 Summary of the image datasets used in our experiments. TCGA, UHCW, CI denote The Cancer Genome Atlas, University Hospitals Coventry and Warwickshire, and the Curie Institute, respectively. Numbers for the updated MoNuSeg dataset are shown in parentheses.

Dataset details	**MoNuSeg**	**CoNSeP**	**TNBC**
Total nuclei	21,623 (28,846)	24,319	4,056
Cancer types	7 (9)	1	1
Images	30 (44)	41	50
Image size	1000 × 1000	1000 × 1000	512 × 512
Source	TCGA	UHCW	CI

significant difference in nuclei densities between the TNBC images and MoNuSeg or CoNSeP images, we set the proportion range $(p_{min}, p_{max}) = (0.5, 1.0)$ for TNBC and $(p_{min}, p_{max}) = (0.01, 0.1)$ for the other two datasets.

4.5.3 Evaluation metrics

In this work, we used two common metrics, namely Average Dice Coefficient (ADC) and Aggregated Jaccard Index (AJI), to evaluate the segmentation performance. The ADC is a pixel-level metric calculated by averaging the Dice coefficient between the segmentation result P and its corresponding ground-truth Q, defined as:

$$Dice(P, Q) = 2\frac{|P.Q|}{|P| + |Q|} \tag{4.8}$$

This means that the Dice coefficient is directly proportional to the number of correctly predicted nuclei pixels divided by the total number of nuclei pixels in both predicted mask and ground truth mask. Note that the Dice coefficient is also closely related to the widely known Jaccard index [130].

The AJI index [130] is an instance-level measurement that has been used in several challenges and recent works. This index is calculated by an iterative algorithm that takes into account both the unmatched detected nuclei's pixels (false positives) and unmatched annotated nuclei's pixels (false negatives).

4.6 Results

4.6.1 Comparison with state of the art methods

Figure 4.3 displays segmentation outputs of sample test images from each dataset. We compared the performance of our unsupervised method with other state-of-the-art supervised and unsupervised methods in nucleus segmentation on the selected public datasets in Table A.3. Note that the supervised approaches all need to split the datasets into training, validation, and test subsets but they don't necessarily provide details about which images were chosen for testing. Therefore, we record our metrics on *all* the images in each dataset. Using significantly more images over which to average the performance, we believe that it ensures a reliable evaluation of segmentation quality.

As can be seen in Table A.3, standard image processing methods (Otsu thresholding [142], Watershed [143], Fiji [108] and CellProfiler [108]), produce inferior results on all three dataset. The self-supervised method presented in [108] provides the possibility of nuclei segmentation without annotations during model training. However, that method still requires a labeled validation set to determine the optimal hyper-parameters for the model. More importantly, it relies on extra image data beyond the provided dataset to boost the performance (see the "Self-supervised" and "Self-supervised + extra WSI" rows in Table A.3). Our method still surpasses their results by a large margin on the two datasets CoNSeP and TNBC. On the MoNuSeg dataset, we obtain superior results compared to the method in [108] when it is limited to using the published dataset.

We also reported the results from recent supervised DL methods and observed that our approach can achieve superior performance than CNN2 [130]. Moreover, the FCN8 and SegNet [118] models perform worse at the object level (lower AJI score) than our model, though they have excellent pixel-level segmentation. Although we record lower performance than the other supervised methods (U-Net and Hover-Net [118]), our model offers two possibilities to enhance the segmentation quality. First, it enables adjusting segmentation output at inference time for each target image thanks to the controlling parameter α (see Section 4.6.3). (The performance metrics recorded in Tables A.3 and 4.3 uses a default value of 0.8 for the decision weight α.) Second, our model is quite modular and can integrate a pre-trained encoder-decoder model using transfer learning or continual learning, whose effectiveness has been demonstrated in many published works.

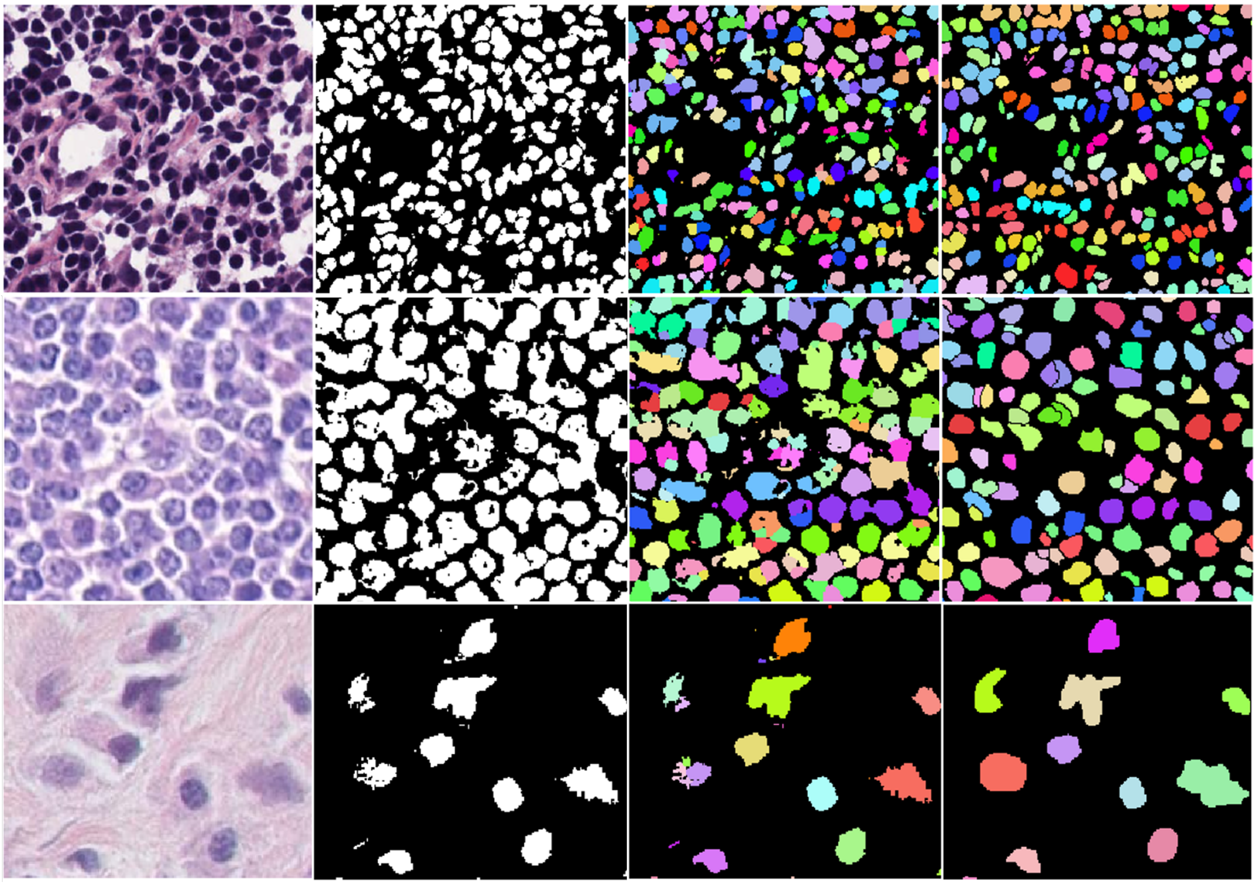


Figure 4.3 Segmentation results on a sample image from each dataset: MoNuSeg (top row), TNBC (middle row) and CoNSeP (bottom row). In each row, from left to right: input image; segmentation map before applying watershed; final segmentation map after applying watershed; ground-truth. In columns 3 and 4, the different colors of nuclei highlight individual instances.

4.6.2 Results on breast cancer images

In the previous section, experiments on different images capturing multiple types of cancer show that our unsupervised method is a competitive approach for nuclei segmentation. As one of our research team's main interests is the analysis of breast tumor cell samples, we examined our segmentation performance on breast cancer images in comparison with a recent work [119] that used various U-Net-style architectures of different complexities. In particular, the encoder portion of the original U-Net network was replaced by various deep convolutional neural networks (DCNNs) of much greater size. The DCNNs employed were the well known VGG-16 [151], ResNet-152 [112] and Inception-v3 [152], all pre-trained on the ImageNet dataset [138]. In addition, an even more complex model, named U-Net Ensem-

Table 4.2 Comparison of different methods on benchmark datasets based on ADC and AJI metrics. Methods marked with * are supervised. Results for the updated MoNuSeg are shown in parentheses.

Methods	**MoNuSeg (updated)**		**CoNSeP**		**TNBC**	
	ADC	AJI	ADC	AJI	ADC	AJI
Otsu [142]	0.0569	0.0032	0.0588	0.0038	0.0047	0.0018
Watershed [143]	0.0562	0.012	0.0582	0.0148	0.0045	0.0092
Fiji [108]	0.6493	0.2733	-	-	-	-
CellProfiler [108]	0.5974	0.1232	0.434	0.202	0.416	0.208
Self-supervised [108]	0.6209	0.3025	-	-	-	-
Self-supervised + extra WSI [108]	0.7477	0.5354	0.587	0.1980	0.5139	0.2656
CNN2* [130]	0.6928	0.3482	-	-	-	-
FCN8* [118]	0.797	0.281	0.756	0.123	-	-
SegNet* [118]	0.811	0.377	0.796	0.194	-	-
UNet* [118]	0.758	0.556	0.585	0.363	0.681	0.514
HoVer-Net* [118]	0.826	0.618	0.664	0.404	0.749	0.590
Our method	**0.7190** (**0.7019**)	**0.3791** (**0.3810**)	**0.5907**	**0.2392**	**0.6662**	**0.3432**

ble, computed the pixel-wise average of three probability maps predicted by three modified U-Nets using pre-trained VGG-19 [151], ResNet-101 [112] and DenseNet-121 [153] networks as encoders. To train these supervised models, the authors of [119] used normalized color images from all tissue types in the updated MoNuSeg dataset, excluding the breast. After training, the models were evaluated on the excluded breast images of the MoNuSeg dataset (8 images, which we refer to as MoNuSeg Breast), and on the TNBC dataset (50 images). Table 4.3 shows the recorded results of these different approaches. Once again, the classical unsupervised algorithms perform poorly on the target breast cancer images from both datasets. Meanwhile, the U-Net variants produce significantly better segmentations. Compared with these methods, our model outperforms all of them at the pixel level (ADC), but does not reach their instance-level scores (AJI). Another point worth mentioning is that with these supervised U-Net models, there is a significant drop in segmentation performance from

MoNuSeg to TNBC, especially the ADC score, although they capture the same type of breast tissue. This demonstrates the limited generalization capability of supervised DL networks compared to our method.

4.6.3 Ablation study

We investigated the contribution of different elements of our proposed framework by configuring the two principal parameters. The first one is the hyper-parameter λ used in the weighted cost function (Equation 4.6) to optimize the encoder-decoder network. The second one is the controlling parameter α used in the decision block at the inference phase (Equation 4.1).

Impact of lambda value

During training, the encoder-decoder network learns from two data sources: the synthesized dataset and the approximate ground-truth obtained by thresholding the nuclei attention map of the target image. This learning process is controlled by the hyper-parameter λ in Equation 4.6. Note that when $\lambda = 0$, the network merely learns from the synthesized dataset. The impact of different λ values is shown in Figure 4.4. Here, we averaged the output scores of all the images across all three datasets. We also canceled the effect of the decision block at the inference stage by setting $\alpha = 0$. Compared with the case when $\lambda = 0$, both the ADC and AJI scores increase slowly as λ increases within the range $\in [0.002, 0.01]$. Moreover, there is no stochastic drop in performance when λ changes from zero to a positive value. This demonstrates that λ's effect on model performance is reliable.

Impact of decision weight

As in Equation 4.1, the output feature map of the encoder-decoder network is combined with the nuclei attention map of the target image under controlling parameter $\alpha \in [0.0, 1.0]$ in the decision block. Note that $\alpha = 0.0$ and $\alpha = 1.0$ correspond to the extreme cases where only the encoder-decoder block or only the nuclei attention map is used, respectively. In our experiments, we set the default value to $\alpha = 0.8$, but it can be adjusted by the user to adapt to each target image. Figure 4.5a demonstrates that in general (across all images collected from all datasets), a hybrid approach (where $\alpha \in [0.2, 0.8]$) is beneficial compared to using only the encoder-decoder block or the nuclei attention pathway. This effect is also observed in Figures 4.5c and d for the CoNSeP and TNBC datasets, respectively. However, the trend is not identical for the MoNuSeg dataset as shown in Figure 4.5b. Although the

Table 4.3 Comparison of different methods on Breast cancer type based on ADC and AJI metrics. Methods marked with * are supervised.

Methods	**MoNuSeg Breast**		**TNBC**	
	ADC	AJI	ADC	AJI
Otsu [119]	0.1619	0.0456	0.0047	0.0018
Watershed [119]	0.2743	0.0828	0.0045	0.0092
Fiji [119]	0.4411	0.3396	-	-
U-Net(VGG-16)* [119]	0.6511	0.4925	0.5042	0.3538
U-Net(ResNet-152)* [119]	0.6706	0.4396	0.5874	0.4063
U-Net(Inception-v3)* [119]	0.6422	0.4440	0.4703	0.3817
U-Net Ensemble * [119]	0.6957	0.4926	0.6068	0.4836
Our method	**0.7081**	**0.3910**	**0.6662**	**0.3432**

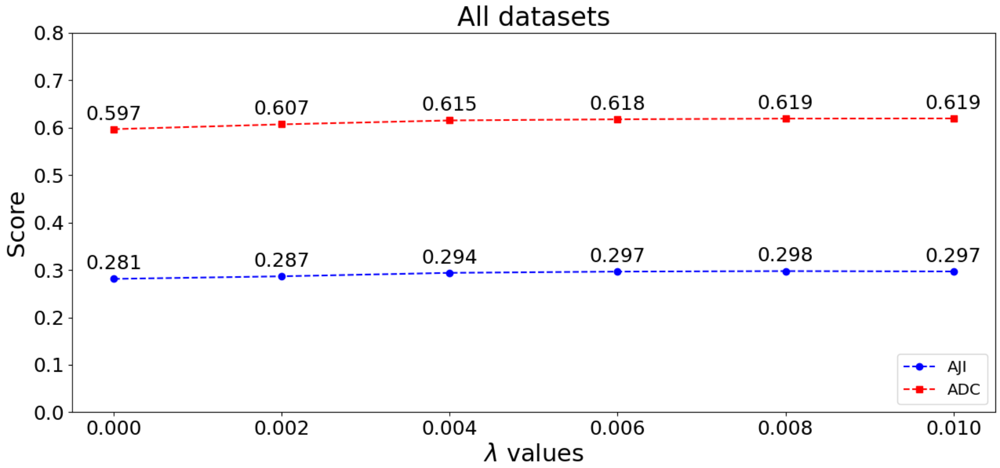


Figure 4.4 Impact of lambda value on segmentation performance.

hybrid option yields better results than using only the DL pathway, using only the nuclei attention branch ($\alpha = 1$) appears to be the best option. This demonstrates the usefulness of making α available for the user to adjust. Note that Figure 4.5 plots the scores averaged across all images in the datasets. This means that for an individual input image, there is potentially an α value more relevant than the default.

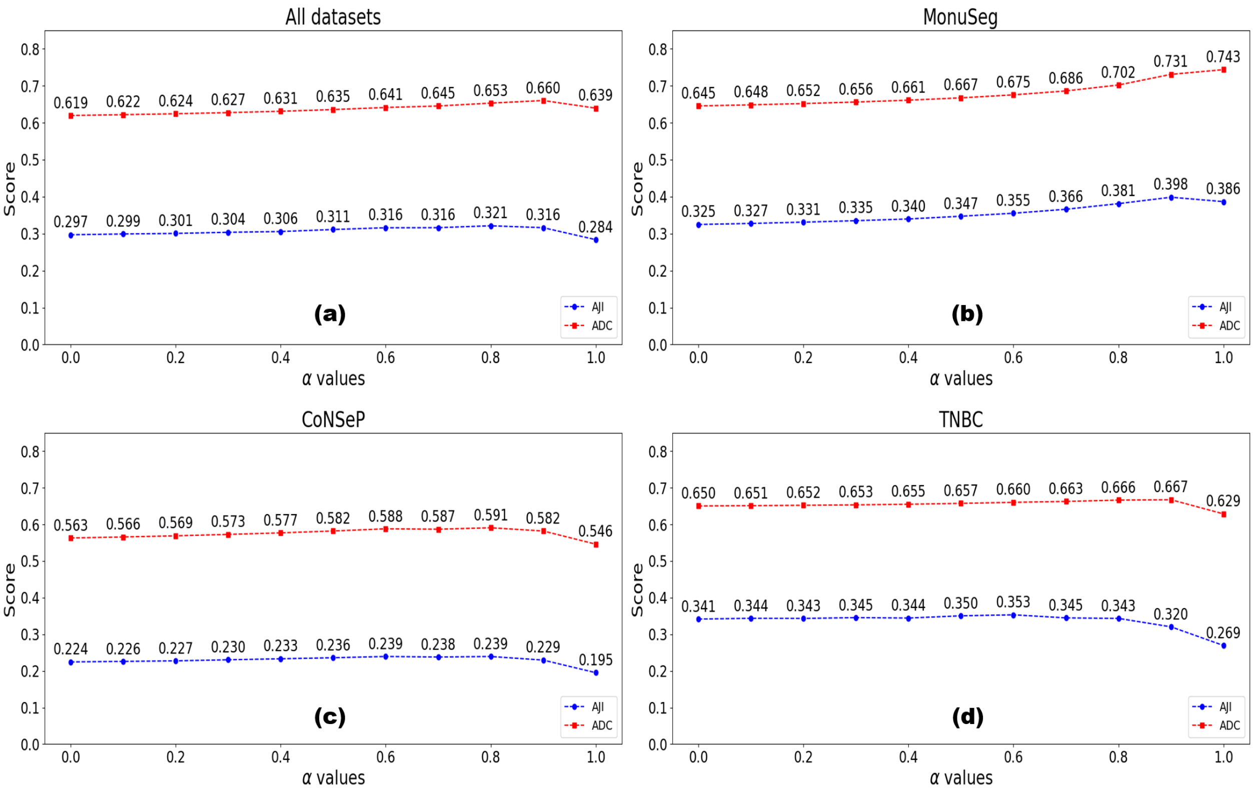


Figure 4.5 Impact of decision weight α on segmentation performance. (a) Average over all datasets; (b-d) Results for MoNoSeg, CoNSeP and TNBC datasets.

4.7 Discussion

This work proposes several contributions toward interpretable and domain-adaptable model design. Deep learning models learn features automatically from raw images and they are assumed to be able to explore hidden image features that machine learning engineers were unaware of [154]. Hence, DL models are often treated as black boxes. There is a growing body of work seeking to understand the predictions of these models using explainable AI approaches, notably saliency maps. Unfortunately, these maps are not very transparent because they are based on features that cannot be easily interpreted or have physical meaning. Ensemble learning approaches make the task of interpreting saliency maps even more challenging. This is because each sub-model works on its own hidden feature space derived from the raw input image, and their saliency maps are often not matched with each other. In contrast, we propose strategies to actively develop an interpretable model by imposing certain parameter constraints related to histological image acquisition. Moreover, we provide a user-adjustable mechanism in the decision block to separate the contributions of different components in our framework.

Most of the current works use supervised learning on a benchmark dataset. In reality, the experimental conditions influencing the target data can change frequently, depending on many factors. As a result, after such a DL model is designed and validated on a benchmark, it is the responsibility of histopathologists to create their specific datasets and optimize the designed model, a task that can prove difficult. The main disadvantage of domain adaptation and self-supervised methods is that they both assume that target data, even without annotations, are accessible during training. This assumption is often invalid in histopathological settings because images in target domains are generally unknown before training the model. Furthermore, as domain shift can occur between different patients, it becomes necessary to collect future patient data in advance, but this is an impractical task [89, 140]. The objective of our work is to propose a method that can help pathologists to benefit directly from the deployed architecture because: (1) it does not require collecting future images and annotating them; (2) information from every target data sample is exploited for adaptation purposes before producing the final segmentation map.

We have shown that our proposed model achieves top performance among unsupervised algorithms and is a potential approach to compete with supervised methods. Its major limitation is the low instance segmentation scores; this is dependent upon the watershed algorithm used to produce the final segmentation map. A deeper encoder-decoder network fed with higher resolution images could mitigate this problem, although this would impose higher computational complexity and longer inference times. Note that we down-scaled the original images several times and this worsens the touching nuclei phenomenon.

4.8 Conclusion

In this study, we presented a novel unsupervised learning model for the segmentation of nuclei in *H&E* stained images. Its benefits of interpretability and domain-adaptability make the design suitable for deployment in a variety of histopathology settings. Our future work will investigate how reinforcement learning assisted by pathologists could be integrated effectively into the model. We will also assess how the performance can be improved by continual learning on the encoder-decoder network, i.e. by fine-tuning a network previously trained on patients' samples acquired in a given laboratory setting. These developments will allow pathologists to have more confidence when using our automatic segmentation algorithm.

4.9 Acknowledgment

This work was funded by the Quebec's consortium for research and innovation in medical technologies (MEDTEQ), the TransMedTech Institute and the Oncopole. The authors would like to thank Dr. Roula Albadine and Cécile Darviot for insightful discussion. We thank Philippe Debanné for useful comments on this manuscript.

CHAPTER 5 ARTICLE 2: MULTI-DOMAIN LEARNING CNN MODEL FOR MICROSCOPY IMAGE CLASSIFICATION

Duc Hoa Tran[1], Michel Meunier[2], Farida Cheriet[1]

[1]Department of Computer and Software Engineering, Polytechnique Montréal, Canada
[2]Department of Engineering Physics, Polytechnique Montréal, Canada

Presentation This work was submitted to the journal "Computer Methods and Programs in Biomedicine Update" on May 30, 2021.

5.1 Abstract

For any type of microscopy image, getting a deep learning model to work well requires considerable effort to select a suitable architecture and time to train it. As there is a wide range of microscopes and experimental setups, designing a single model that can apply to multiple imaging domains, instead of having multiple per-domain models, becomes more essential. This task is challenging and somehow overlooked in the literature. In this paper, we present a multi-domain learning architecture for the classification of microscopy images that differ significantly in types and contents. Unlike previous methods that are computationally intensive, we have developed a compact model, called Mobincep, by combining the simple but effective techniques of depth-wise separable convolution and the inception module. We also introduce a new optimization technique to regulate the latent feature space during training to improve the network's performance. We evaluated our model on three different public datasets and compared its performance in single-domain and multiple-domain learning modes. The proposed classifier surpasses state-of-the-art results and is robust for limited labeled data. Moreover, it helps to eliminate the effort for designing a new network when switching to new experiments.

5.2 Introduction

There exists a wide range of microscopy assays to reveal complex properties of cellular structures (tissues, cells, or subcellular components) and each set of images produced in a laboratory typically forms a different visual domain. Although Deep Learning (DL) models could yield excellent classification performance, they are highly specialized to each domain [155, 156]. At the same time, designing and training an appropriate deep model are relatively complex operations to carry out successfully, even for experienced scientists [157]. Thus, there is growing interest in developing a single model that can be deployed for various biomedical studies without adjusting its parameters. This is a challenging task as it requires deep learning models to learn a unified feature representation across different domains.

To contribute to this research effort, we aimed at designing a deep convolution neural network (CNN) to learn unified representations for the classification of microscopy image sets that have significantly different characteristics. This problem belongs to Multi-Domain Learning (MDL) and can be distinguished from the related domain adaptation technique in two ways: the domain shift and the learning sequence. First, the domain shift refers to the visual difference between image domains, including image content (objects of interest) and image appearance (style). While standard domain adaptation methods deal with the change in style and not the objects of interest, our model handles both changes in image content and style. Second, in terms of the learning sequence, typical domain adaptation approaches learn multiple domains sequentially to maximize their performance in a target domain. However, after adapting from a source domain to the target one, the model cannot maintain its initial performance on the source domain or it cannot learn without forgetting [155]. In this sense, domain adaptation is like transfer learning, where DL models are trained on a common large dataset, before being fine-tuned on the domain of interest. By contrast, our proposed model learns multiple domains simultaneously and aims at achieving high performance on all the learned domains.

In this work, we design a deep CNN architecture that combines an inception module and depth-wise separable convolution layers. These two techniques, introduced in GoogleNet [158] and [92], are popular in the design of many deep neural networks. However, to the best of our knowledge, this is the first work that explores their combination for multi-domain learning of microscopy images obtained from different imaging devices and with different objects of interest. The proposed model is lightweight and scalable. In addition, we introduce an optimization approach for feature regularization during training to enhance the network's performance, allowing it to beat the state-of-the-art models.

To sum up, the major contributions of this study are:

- We propose an MDL model for learning simultaneously different microscopy image domains. It could work effectively in different applications without requiring the adaptation of domain-specific parameters.
- We formulate a simple yet effective optimization function to regulate feature space, improving network performance.
- Our proposed model is remarkably compact and robust against limited available train-

ing data and outperforms the best results published as yet on three public datasets.

5.3 Related Works

DL algorithms or deep neural networks have emerged as the dominant methods in every application of biomedical image analysis, including microscopy [4], [159]. Instead of using handcrafted feature extractors as in conventional machine learning methods, CNN models learn by themselves to extract the optimal features from input images. However, it is still challenging to design a model that can extract a unified feature representation from multiple microscopy image domains because of the highly specialized experiments. In the literature, most DL-based analyses have used transfer learning, as it can produce significantly better results than when training from scratch. Comprehensive reviews of different applications using transfer learning can be found in [4], [159]. In this approach, a DL model is pre-trained on a large dataset of labeled natural images like ImageNet and then fine-tuned on a target datasets that usually has a few labeled images.

DL-based domain adaptation approaches have also been investigated for digital pathology [109, 160, 161]. In these approaches, the feature extractors of a DL model are first trained on a source domain and then adapted to the target domain via a retraining process. The two domains are supposed to be similar or undergo a minor domain shift, in the sense that the image style or appearance changes but the image content or objects of interest are the same. The domain adaptation methods conventionally tackle the problem by normalizing the imaging parameters, such as staining normalization, or aligning the source representation with the target one using feature space transforms [162, 163]. Like in transfer learning, during fine-tuning of the DL network, the pre-trained parameters are specifically adjusted for the target domain and thus, the fine-tuned model cannot be reused on the source dataset. Many of the latest techniques use Generative Adversarial Networks (GAN) to learn domain-invariant features [164]. For example, the authors of [165] propose to use adversarial training to align the feature distributions of shifted domains in the classification of prostate cancer images acquired from different scanners.

Recently, MDL has become a popular topic in computer vision, but there are still a few published articles for biomedical image analysis, especially for microscopy images. The multi-domain adversarial learning approach presented in [163] was the first work using MDL in bio-image informatics. The authors experimented on cell-level fluorescence images obtained from three different centers and used a pre-trained VGG-16 [166] network as the feature extractor to feed the cells classifier. Nevertheless, that study only considered the variation in image appearance, while the type of content (i.e cells) remained unchanged.

5.4 Materials and Methods

5.4.1 Convolutional neural network architecture

In this section, we present a deep multi-domain CNN model, named Mobincep, which is compact but offers powerful classification capability for various types of microscopy images. To achieve efficient feature extraction, the network's design is based on the combination of the inception structure [158] and depth-wise separable convolution [92]. Also, we describe a relevant training strategy, particularly the formulation of an integrated loss function for network optimization. The following sections describe the construction of our model.

Depth-wise separable convolution layers

The depth-wise separable convolution effectively reduces the computation complexity of the standard convolution by dividing the calculation into two separate and consecutive steps: depth-wise and point-wise (or 1×1) convolution [167]. In the first step, each input channel is convolved with kernels that have only a single channel. Then, point-wise convolution creates a linear transformation of the corresponding output values across channels. We illustrate the two convolution approaches in Figure 5.1. As depth-wise convolution reduces the number of deep CNN network parameters significantly, it also helps to decrease the possibility of over-fitting to a specific image dataset.

Compared with standard convolution, the use of depth-wise separable convolution produces more discriminative features due to the decoupling of cross-channel and spatial correlations as suggested in [168]. It can also help to promote the performance for learning natural images from multiple visual domains when replacing standard convolutions in a pre-trained ResNet-26 [167]. Based on this observation, we exploit depth-wise separable convolution layers as one of the main strategies to design a compact model for microscopy images in multiple imaging settings.

Inception module

We illustrate the layout of an inception module in Figure 5.2. It is a set of multiple convolution branches having different kernel sizes, where the output feature maps from each branch are concatenated and used as the input for the subsequent layer [158]. In this configuration, the use of the point-wise (1×1) convolution and average pooling layer in each branch reduces the dimensions of the feature maps, leading to a remarkable reduction of multiplication operations. Meanwhile, the combination of various convolution kernel sizes helps to detect

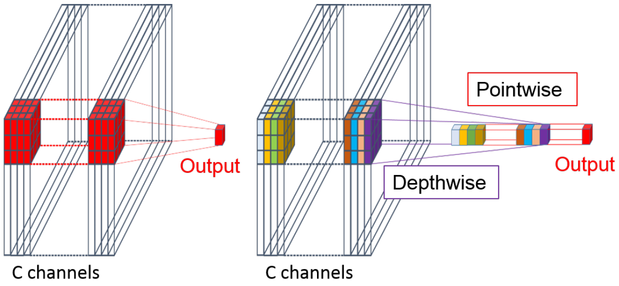


Figure 5.1 Example of depth-wise separable convolution compared with standard convolution operations. In the standard convolution (on the left), the multi-channel input is convolved with kernels having smaller spatial dimensions (3×3) but the same number of channels (C). The depth-wise separable convolution (on the right) divides the calculation into two separate and consecutive steps: depth-wise (convolution with 3×3 kernel of 1 channel) and point-wise (convolution with 1×1 kernel of C channels).

features at different scales. This is advantageous for microscopy images that typically express a wide variety of object morphologies and sizes.

Proposed network architecture

The overall structure of our proposed Mobincep network is shown in Fig. 5.3. Through a buffer convolution layer, the raw input images are fetched to the inception module, comprising four branches with different kernel sizes. In this work, we choose 1×1, 3×3, and 5×5 filters as they have proved to be effective feature extractors. The concatenated feature maps are then delivered to a stack of multiple depth-wise separable convolution layers. Each of these layers is followed with Batch Normalization (BN) [95] and in-place Rectified Linear Unit (ReLU) [169] activation functions. The role of BN is to normalize each layer's output

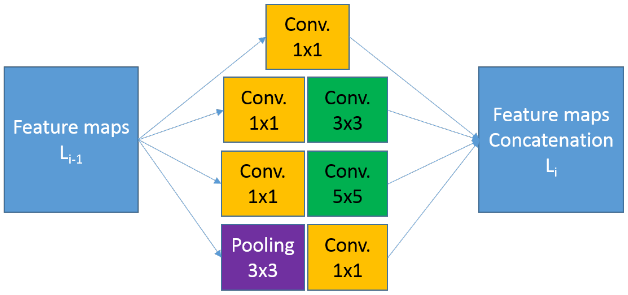


Figure 5.2 Design of the inception module. The feature maps of previous layer L_{i-1} are processed by multiple branches with different reception fields. The outputs from each branch are concatenated into a set of feature maps L_i to be used as the input to the following layer.

such that it has zero mean and unit standard deviation, leading to faster convergence and bypassing local minima. For its part, the ReLU activation function has properties that help feed-forward neural networks to optimize easily with gradient-based convergence and generalize well on various data domains [7]. Lastly, the average pooling layer compresses the extracted feature maps into a feature vector. At the output, a linear layer combines the vector elements to produce the prediction probability for every image class. The class that has the highest probability from the output layer will be selected as the predicted class for the input image.

5.4.2 Network optimization

We propose a new model optimization approach to regularize the extracted feature space of microscopy images during training. The conventional loss function used in the optimization of deep CNN classifiers may not effectively regulate the latent space, leading to low performance and applicability to different sets of data. Therefore, we formulate two additional loss terms to encourage the feature representation of samples within each class to converge to a compact corresponding cluster. Assuming there are K categories to be classified, the integrated cost function for training the Mobincep network is expressed as:

$$L = L_{CE} + \gamma_1 \frac{1}{\Sigma_{k=1}^{K}(d^2(\mu_k, \mu))} + \gamma_2 \Sigma_{k=1}^{K}(s_k^2) \tag{5.1}$$

In equation (5.1), L_{CE} is the conventional cross-entropy loss criterion for classification, which is calculated as:

$$L_{CE} = -\frac{1}{N_B} \Sigma_{i=1}^{B} \Sigma_{c=1}^{K} I_{i,c} \log \frac{\exp(y_{i,c})}{\Sigma_{c=1}^{K} \exp(y_{i,c})} \tag{5.2}$$

where N_B is the training batchsize; $I_{i,c} = 1$ if label c is the correct classification for image sample i and $I_{i,c} = 0$ otherwise; $y_{i,c}$ is the raw output probability of the network for the sample i to have class label c. Also in equation 5.1, we add the two other loss terms to impose an additional constraint for the training. In the first term, $d(\mu_k, \mu)$ denotes the distance between the centroid of each cluster in the feature space and the centroid of all latent points. In the second term, variable s_k represents the scattering of each cluster, calculated as the sum of all distances between each latent point and the centroid μ_k of its cluster. Variables γ_1 and γ_2 are the weights to balance the three loss terms in the total cost function. At each training iteration, the cluster centroids in the latent space are quickly determined by using conventional K-Means clustering algorithm.

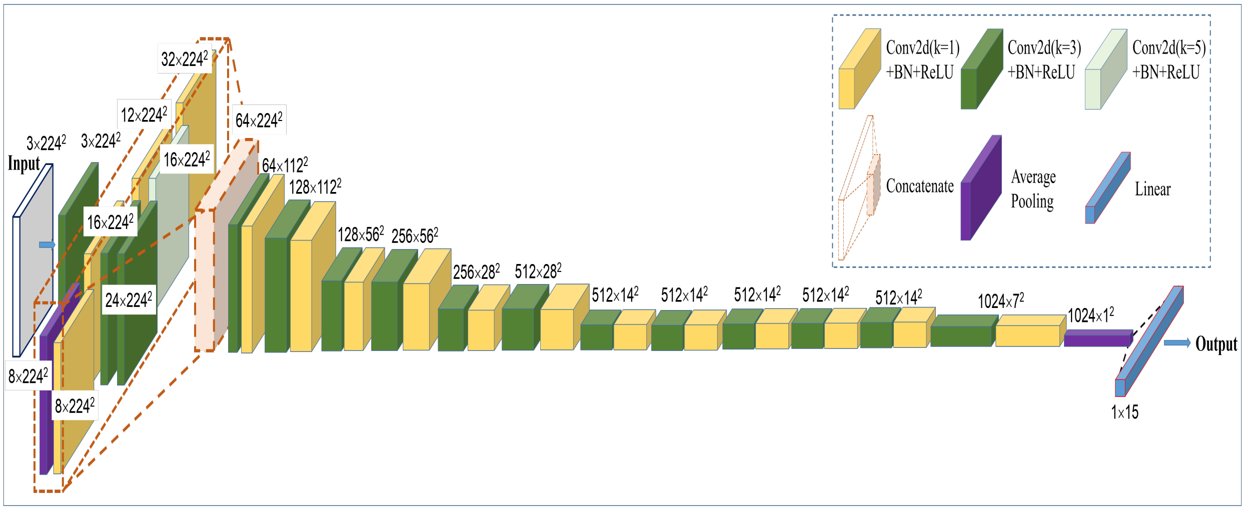


Figure 5.3 Architecture of the proposed Mobincep model. Each of the convolution layers (Conv2d) is followed by a Batch Normalization layer (BN) and in-place Rectified Linear Unit (ReLU) activation function. The raw input images are passed through a buffer convolution layer to the inception module, which comprises four branches with different kernel sizes. Then, the concatenated feature maps are forwarded through a stack of multiple depth-wise separable convolution layers. After that, the extracted features are reduced in spatial size by average pooling operation and linearly combined to produce the output prediction.

Intuitively, when minimizing this loss function, the addition of two new loss terms helps in two ways: decrease the scattering (i.e. the embedded distances) of the input samples around their centroids and increase the embedded distance between clusters. This helps to better discriminate between clusters or classes.

5.4.3 Experimental setup

In this section, we describe our experiments with the Mobincep model on datasets from three different imaging domains.

Datasets

To create a dataset (Mix) that represents different imaging domains, we use three public microscopy datasets: Lymphoma (Lym), composed of tissue sample images [170]; Pap-smear (Pap), with images of cells [171]; and HeLa, with images of sub-cellular organelles [170]. The characteristics of these datasets are summarized in Table A.1. Example images from each dataset are shown in Figures 5.4, 5.5 and 5.6.

Network training, validation and testing

As shown in Table A.1, the experimented datasets vary significantly in terms of image characteristics. For the model to handle them properly, we pre-processed the raw images by resizing them to $224 \times 224 \times 3$ and normalizing them to have the same dynamic intensity range. During training, we applied online data augmentation, combining a wide range of common image transformations, including rotation, flipping, cropping, and affine transformations (translate, scale, shear). We process the input images in batches of four. The network layers were initialized with the Kaiming uniform method [172]. We used the modified Adam optimizer,

Table 5.1 Microscopy image datasets used experimentally

Dataset	**# Images**	**# Classes**	**Dimensions**
Lymphoma	375	3	$1388 \times 1040 \times 3$
Pap-smear	917	2	$45 \times 43 \times 3$ to $768 \times 284 \times 3$
HeLa	862	10	$382 \times 382 \times 1$
Mix	2154	15	$45 \times 43 \times 3$ to $1388 \times 1040 \times 3$

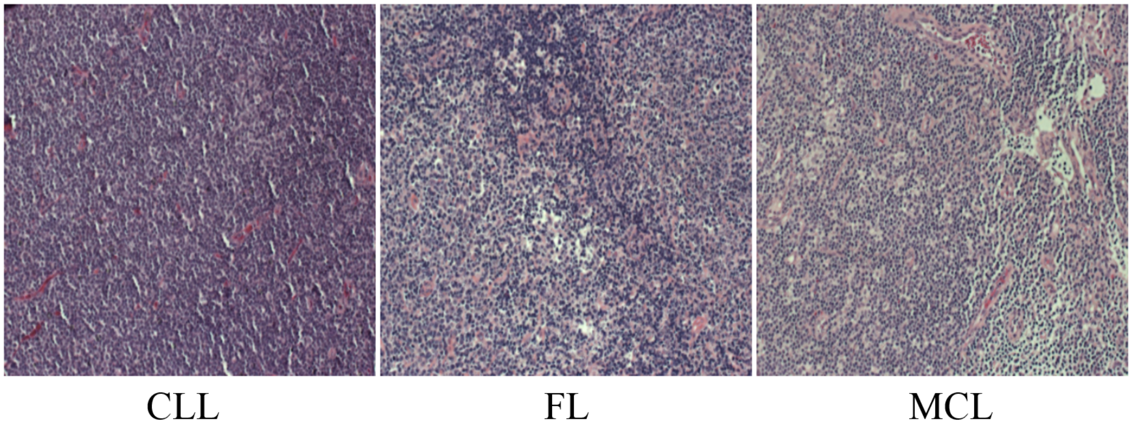


Figure 5.4 Example of Lymphoma images in three different cases. The images were obtained from Hematoxylin- and Eosin- ($H\&E$) stained tissue samples using brightfield microscopy. There are three types of malignant lymphoma, i.e. cancer affecting lymph nodes: chronic lymphocytic leukemia (CLL), follicular lymphoma (FL), and mantle cell lymphoma (MCL).

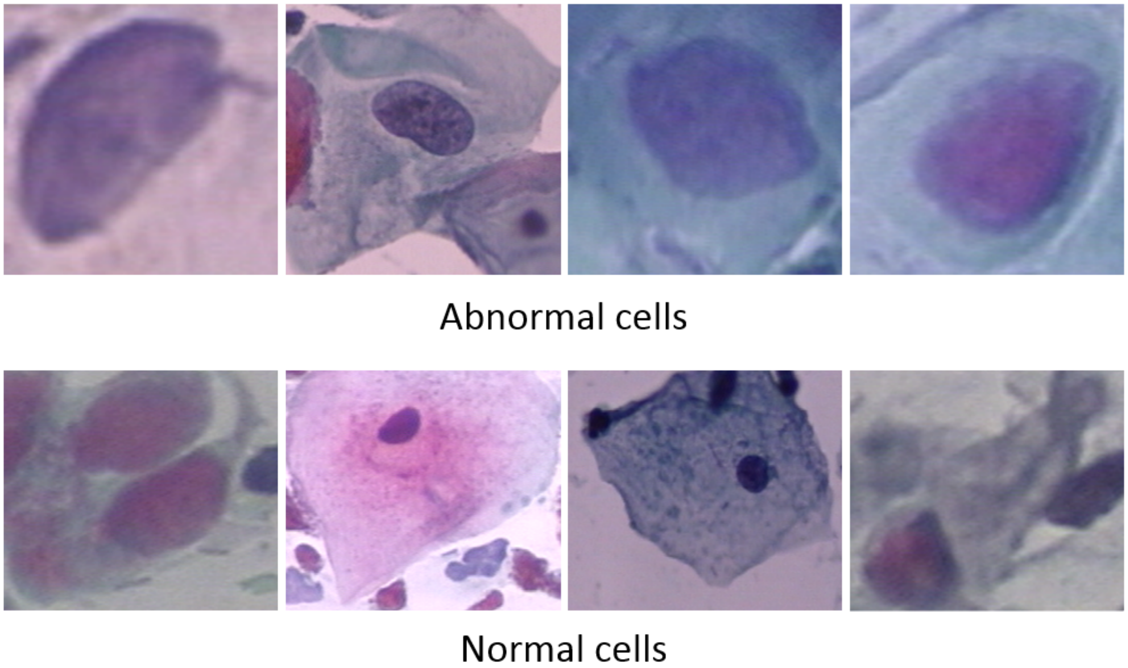


Figure 5.5 Example of Pap-smear images in two cases: Abnormal and Normal cells. In the sample preparation, a specimen of human cells is smeared onto a glass slide and colored using the Papanicolaou method. The abnormal cells are associated with the pre-cancerous stage.

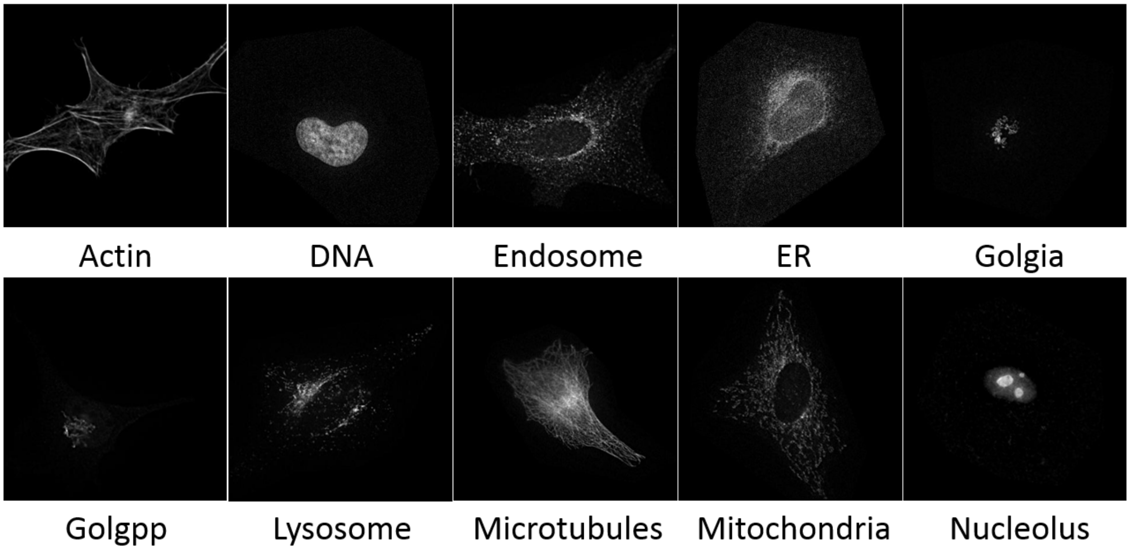


Figure 5.6 Example images of HeLa dataset. It comprises fluorescence microscopy images of sub-cellular organelles in HeLa cells, which are stained with various organelle-specific fluorescent dyes. There are ten categories: Actin, DNA(Nuclei), Endosomes, ER (Endoplasmic reticulum), Golgia (Giantin), Golgpp (GPP130), Lysosome, Microtubules, Mitochondria and Nucleolus.

referred to as AMSGrad [173], with a small learning rate of 10^{-4}. We validated the performance of the model using a 5-fold cross-validation strategy. In each fold, we randomly split each dataset into training, validation, and testing subsets. Specifically, we used 60% of the images for training, 20% for validation, and the remaining 20% to assess network performance. To obtain the model with the lowest validation error, the early stopping strategy was adopted, where the training stops when there is no better model after a certain number of training epochs (or patience period). As the training process includes random processes (for instance, the data augmentation), we ran the experiments five times and recorded the average result.

5.4.4 Model complexity

Figure 5.7 shows the complexity of our proposed model compared to prominent CNN models from the literature. The number of trainable parameters in our Mobincep network is remarkably small compared with most of the recently published networks. For example, it requires about 13 times fewer parameters than Inception-ResNet-v2 [124], which has 56 million parameters. Although the MobileNetV2 [93] network has a slightly smaller number of parameters, our model can outperform MobileNetV2 by a large margin in various experiments, as shown in the following section.

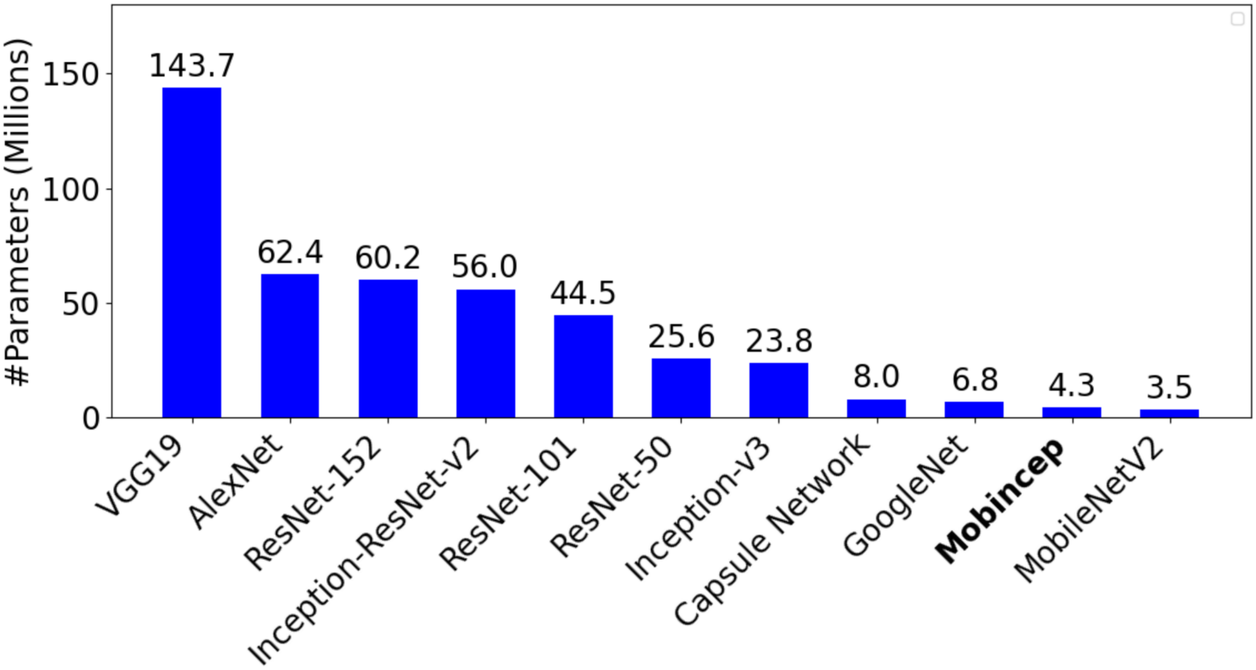


Figure 5.7 Complexity of state-of-the-art CNN models compared by number of trainable parameters.

It took less than 2 hours to train the model with each fold of the dataset. We used a GPU (model NVIDIA Tesla V100 SXM2 with 16 GB memory) to train the network in an end-to-end fashion.

5.5 Results

We performed four main experiments to test the ability of the proposed model in dealing with microscopy images from three different visual domains. In each experiment, we report the results in both the MDL and Single-Domain Learning (SDL) modes.

5.5.1 Analysis of classification results

We compared the classification accuracy of our approach to recent deep CNN models pre-trained on the ImageNet dataset: VGG19 [166], GoogleNet [158], ResNet-101 [122], Inception [174], and Inception-ResNet-v2 [124]. We also trained the lightweight CNN model MobileNetV2 [93] from scratch using the same training conditions as in the original research. This model has very low computation and model complexity and thus we could train it from scratch with our limited training data.

Table 5.2 Top-1 classification accuracy on the different datasets compared to recent deep CNN models.

Model	**Mix**	**Pap**	**Lym**	**HeLa**
VGG19	83.92±2.1	84.32±1.9	83.52±1.7	84.73±2.1
GoogleNet	84.11±2.3	84.71±2.1	83.94±1.9	85.42±2.0
ResNet-101	88.14±2.1	89.02±1.8	88.43±1.6	90.63±1.5
Inception	89.33±2.5	89.73±2.2	88.91±2.1	90.89±2.1
Inc.-ResNet-v2	87.34±2.3	87.44±2.4	86.72±2.3	88.75±2.2
MobileNetV2	91.24±2.4	92.44±2.3	88.09±2.1	91.75±2.2
Mobincep	**94.82±2.1**	**94.86±1.9**	**94.11±1.8**	**95.90±1.8**

Table 5.3 Classification metrics per class for separate (at left) and mixed (at right) datasets.

Dataset: Class	**Precision**	**Sensitivity**	**F1**	**Dataset: Class**	**Precision**	**Sensitivity**	**F1**
Pap: Normal	0.888	0.882	0.885	Mix: Normal	0.906	0.874	0.890
Pap: Abnormal	0.956	0.960	0.958	Mix: Abnormal	0.954	0.966	0.960
Lym: MCL	0.952	0.886	0.918	Mix: MCL	0.966	0.888	0.925
Lym: FL	0.952	0.984	0.968	Mix: FL	0.940	0.962	0.951
Lym: CLL	0.926	0.946	0.936	Mix: CLL	0.908	0.956	0.931
Hela: Nucleolus	0.988	1.000	0.994	Mix: Nucleolus	0.976	0.988	0.982
Hela: Mitochondria	0.956	0.894	0.924	Mix: Mitochondria	0.904	0.948	0.925
Hela: Microtubules	0.952	0.988	0.970	Mix: Microtubules	0.966	0.954	0.960
Hela: Lysosome	0.944	0.928	0.936	Mix: Lysosome	0.928	0.940	0.934
Hela: Golgpp	0.946	0.904	0.925	Mix: Golgpp	0.974	0.846	0.905
Hela: Golgia	0.912	0.940	0.926	Mix: Golgia	0.876	0.966	0.919
Hela: ER	0.958	0.952	0.955	Mix: ER	0.932	0.964	0.948
Hela: Endosome	0.928	0.914	0.921	Mix: Endosome	0.926	0.838	0.880
Hela: DNA	0.978	1.000	0.989	Mix: DNA	0.978	1.000	0.989
Hela: Actin	1.000	1.000	1.000	Mix: Actin	0.990	1.000	0.995

As shown in Table 5.2, our approach achieves better top-1 accuracy than the pre-trained deep CNN models either when learning images from three domains simultaneously in the Mix dataset or when learning from each of the single domain datasets. Notably, it gains over 10% accuracy compared to the VGG or GoogleNet network. Compared to the trained-from-scratch MobileNetV2, our model also produces better top-1 accuracy across the four datasets. More importantly, our performance results are more consistent on the Lymphoma dataset.

We examined the classification metrics of the proposed model in the SDL and MDL modes, as shown in Table 5.3. When learning on the mixed dataset, the precision and sensitivity of detecting abnormal samples were 95.4% and 96.6%, respectively, while the F1 score was

96.0%. These scores are slightly better than when the network was trained on the Pap-smear dataset alone. Figure 5.8a and 5.8d plots the true-positive rates (TPR) against the false-positive rates (FPR), known as ROC curves, for classifying abnormal samples from normal samples in the SDL and MDL modes, respectively. In the case of MDL, the AUC value is 0.996, which is also better than the single-domain value (0.978).

For the Lymphoma images, the F1 score values were higher than 92.2% for all three classes. We can also notice a significant variation in the precision and sensitivity across the classes, e.g. the low sensitivity for MCL compared with FL and CLL. This fluctuation is similar in both MDL and SDL. In practice, we could improve the classification performance by using a suitable decision threshold. This value could be selected based on the ROC curve for the MCL class in Figure 5.8b, where the AUC of the MCL class is close to the CLL and FL classes.

For the HeLa dataset, the MDL model yielded F1 scores of over 90.4%, except for the Endosome class which has a score of 87.6%. Again, this lower performance resembles the network's output when learning from the single dataset. These results coincide with the fact that experts find it challenging to distinguish between Endosomes and Lysosomes or between Golgpp and Golgia proteins. The ROC curves in Figure 5.8d confirm that the proposed classifier works well in recognizing subcellular organelles in fluorescent images as the AUC values are at least 0.969.

As can also be seen from the figure 5.8, we obtained micro-average and macro-average AUC scores close to 1.0 in all datasets.

5.5.2 Impact of available training data

Next, we investigated the impact of limiting the number of labeled images on the model's performance. For each dataset, we experimented with three different ratios for 5-fold cross-validation by increasing the ratio of training data: 20/20/60, 40/20/40, and 60/20/20. We illustrate the results in Figure 5.9.

For the Pap smear dataset, it needs to learn from around 90 training images per class to distinguish between normal and abnormal classes with an accuracy of 90.74%. This accuracy level increased to 94.86% when the number of images available for training was tripled. On the other hand, the classifier required only about 50 and 20 images per class to reach an accuracy of around 90% on the Lymphoma and HeLa datasets, respectively. By comparison, existing machine learning methods require at least 70 labeled images per class to achieve cross-validation accuracy close to 90% on fluorescence microscopy images like those in the HeLa dataset [175]. In MDL mode on the Mix dataset, we can see that Mobincep reached the 90% accuracy level with 30 images per class. These results not only attest to the classifier's generalization ability on unseen data but also reveal a very useful property of Mobincep, which is to reduce the costly labeling effort needed from experts.

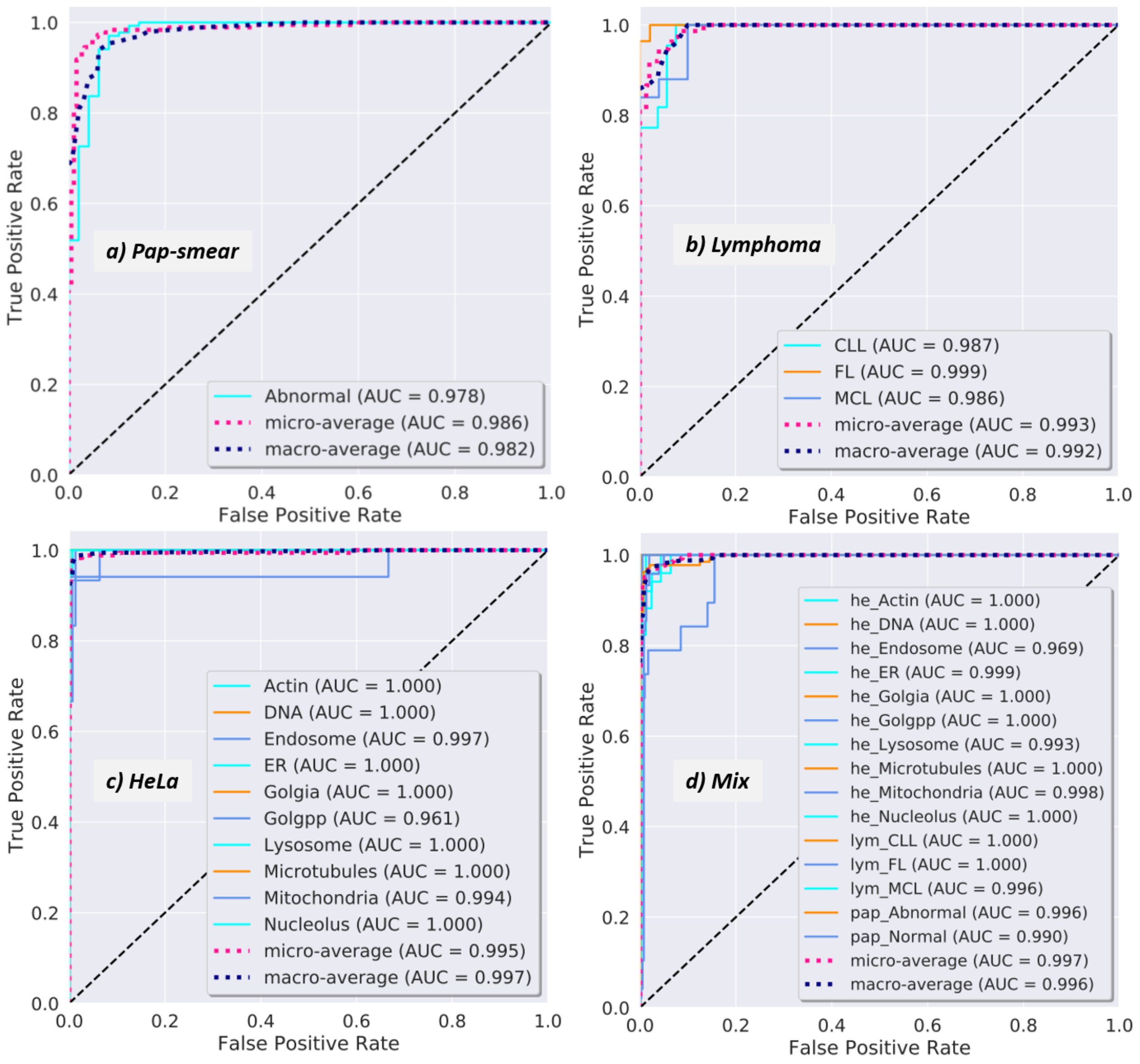


Figure 5.8 ROC curves for the classification task on the different datasets: a) Pap-smear, b) Lymphoma, c) HeLa, d) Mix. The micro-average AUC score is calculated sample-wise, computing average value across all classes weighted by the number of samples in each class, whereas the macro-average score is class-wise, computing unweighted average value across classes.

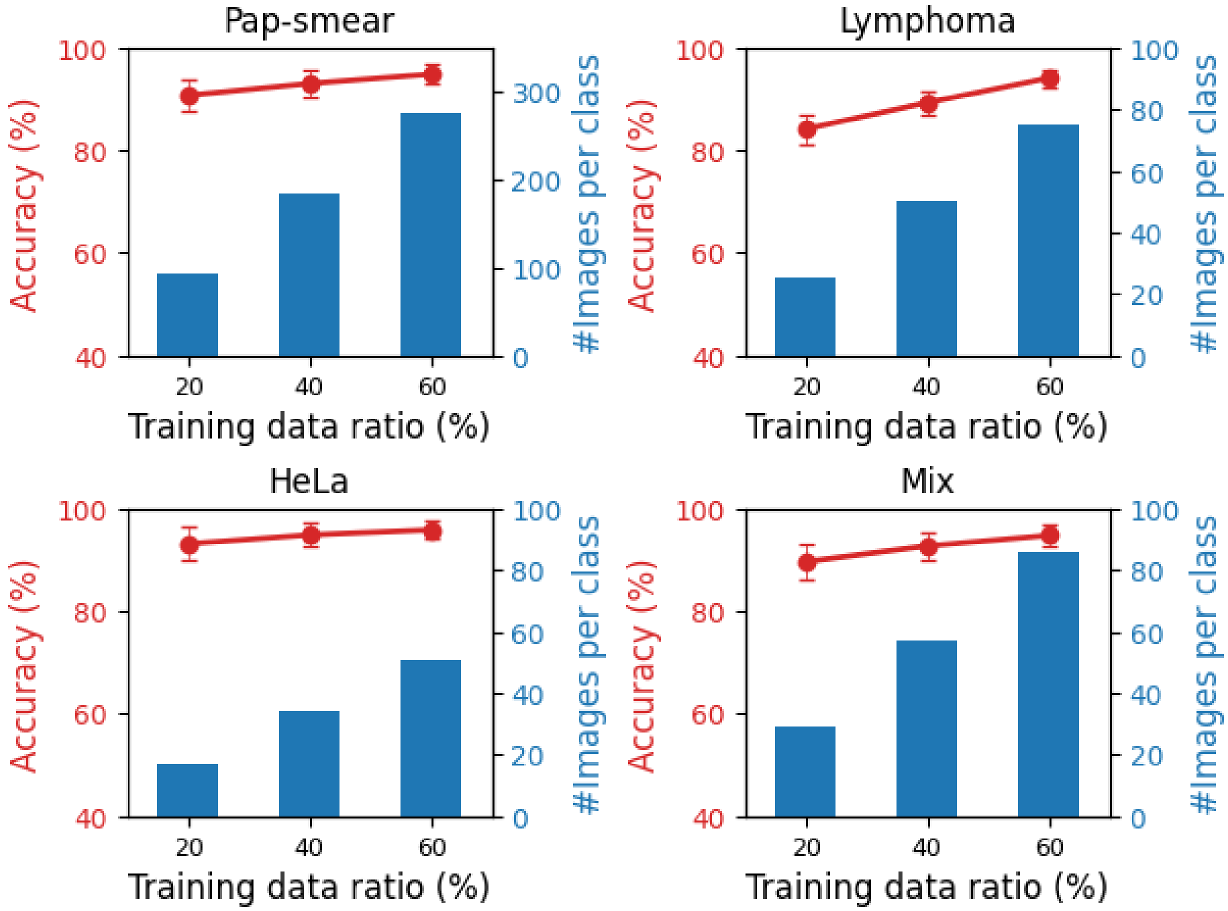


Figure 5.9 Impact of training data volume on classification accuracy.

5.5.3 Impact of the regularization technique

To validate the contribution of the proposed regularization technique, we compared the performance of the network with the baseline case where only the conventional cross-entropy loss function (the L_{CE} term in equation (5.1)) was used during optimization.

As shown in Table 5.4, for all experimented datasets, regulating features during optimization improved the classification accuracy. On the multi-domain Mix dataset, it allowed the model to gain 1.56% in average accuracy.

5.5.4 Comparison with state of the art methods

In Table 5.5, we compare our proposed Mobincep network with recent methods that achieved the highest published results on the three experimented datasets (Pap, Lym, and Hela).
Looking at the Pap smear dataset, all of the methods using a single deep CNN network give no improvement over conventional hand-crafted feature-based methods, reaching accuracy levels lower than 91%. The accuracy increases notably only when multiple deep networks are combined, reaching around 93% as achieved in [79]. In contrast, our approach produces the best performance by training only a lightweight CNN model with much lower complexity, with 94.02% accuracy in MDL mode and 94.86% in SDL mode.
The Lymphoma dataset appears to be more challenging for designing suitable hand-crafted feature descriptors. The best approach which was proposed in [176], with 93.87% accuracy, merged 8 different deep CNN models. However, we show that while our Mobincep network has only 4.3M parameters, it produces an accuracy of more than 94%.
For the HeLa dataset, the Capsule Neural Network is the best option among CNN models, but even its accuracy of 93.08% is still well below the top performance (95.3%) produced by the hand-crafted method described in [62]. Our Mobincep network could surpass this value, reaching 95.9% accuracy whether training on the mixed dataset or the HeLa dataset alone. This confirms yet again the effectiveness of our model and its ability to generalize across different image domains.

Table 5.4 Impact of feature regularization on classification accuracy (%)

Model	**Mix**	**Pap**	**Lym**	**HeLa**
Baseline	93.26±2.6	93.08±2.4	91.23±2.2	93.15±2.1
Mobincep	**94.82±2.1**	**94.86±1.9**	**94.11±1.8**	**95.90±1.8**

In the multi-domain learning setting, the proposed network achieves classification accuracy roughly equivalent to its performance when it is optimized on each domain separately. Indeed, while its accuracy decreased by 0.84% on the Pap-smear classes, the average accuracy was maintained on the HeLa classes and slightly improved on the Lymphoma images.

5.6 Discussion

This study aimed to develop a multi-domain learning model for the classification of microscopy images from different domains. Our principal contribution is the design of a compact CNN model that can be trained from scratch on target domains that have a very limited number of image samples. The proposed multi-domain learning approach can facilitate the choice of an analysis tool and the configuration of its parameters to account for different microscopes, objects of interest, and imaging conditions. This can accelerate the pace of investigation and reduce the required expertise in adapting computer algorithms. From a design perspective, this limits the need to select domain-specific hyper-parameters and eases the training process, as we can train only once the model and run on different experimental domains.

Prior studies typically used very deep CNNs or ensembles of these architecture to achieve high performance in clinical or biomedical applications. Nevertheless, this increases the requirement for computer hardware that is not always readily available. More importantly, their high complexity generally renders such networks selective for a certain image domain. As described in section 5.5.1, the very good classification statistics and high AUC values (> 0.95) on multiple domains demonstrate the benefit of using the Mobincep model as an automatic classifier. Our experiments show that the model yields state-of-the-art performance when learning from multiple microscopy image domains. This can be explained by the fact that the model has very low complexity, thus we can train it from scratch directly on the target data, instead of fine-tuning a pre-trained feature extractor using transfer learning.

5.7 Conclusion

In this work, we presented a lightweight CNN classifier for learning multiple domains of microscopy images. Moreover, we formulated a new optimization function and devised a suitable training strategy allowing our network to outperform state-of-the-art methods. The proposed model performs well in multiple applications of microscopy image classification. Because of its low complexity, the approach becomes more appealing for deployment in clinical and biomedical studies. Further research will focus on developing a domain generalization

Table 5.5 Comparison of Mobincep with competing methods on different datasets. All values are accuracies (%). Those for other methods are as published in the literature, with standard devs. when available.

Model	**Pap**	**Lym**	**HeLa**
Spatial adjacent histogram based on adaptive local binary patterns+SVM [177]	88.03 ± 1.7		90.06 ± 1.5
SVM cascaded with a reject option and subspace analysis [178]	90.96 ± 0.5		92.96 ± 1.3
WND-CHARM based on 1025 content descriptors [179, 180]		85.00	87.00 ± 9.00
CP-CHARM based on 953 content descriptors [179]		66.00 ± 1.0	84.00 ± 0.4
Fusion of multiple handcrafted and deep learned features [181]		90.67	
Multiresolution classification system [62]			95.30
Pretrained ResNet-101 [176]		86.40	
Pretrained ResNet-152 [79]	90.87 ± 1.5		
Pretrained Inception-ResNet-v2 [79]	89.25 ± 2.2		92.00
Pretrained Inception-v3 [79, 176]	89.66 ± 1.9	87.47	
Ensemble of pretrained Inception-v3 and ResNet-152 [79]	92.38 ± 1.3		
Ensemble of pretrained Inception-v3, ResNet-152, Inception-ResNet-v2 [79]	**93.04 ± 1.5**		92.57
Fusion of 8 different deep CNN models [176]		**93.87**	
GoogleNet [78]			92.00
Capsule Neural Network (CapsNet) [182]			93.08
Multi-scale CNN [78]			91.00
Mobincep	**94.86±1.9**	**94.11±1.8**	**95.90±1.8**
Mobincep (Multi-domain learning on Mix dataset)	**94.02±1.9**	**94.20±1.8**	**95.90±1.8**

algorithm, such that the model can work well on new microscopy images that are captured under imaging conditions different from those of training images.

Acknowledgment

This work was funded by the Quebec's consortium for research and innovation in medical technologies (MEDTEQ), the TransMedTech Institute and the Oncopole. The authors would like to thank Philippe Debanné for revising this manuscript.

Data availability

The source code to train the proposed network and datasets related to this article can be accessed at `https://github.com/duchoapoly/mobincep`.

CHAPTER 6 ARTICLE 3: WAVEM-CNN FOR AUTOMATIC RECOGNITION OF SUB-CELLULAR ORGANELLES

Duc Hoa Tran[1], Michel Meunier[2], Farida Cheriet[1]

[1]Department of Computer and Software Engineering, Polytechnique Montréal, Canada
[2]Department of Engineering Physics, Polytechnique Montréal, Canada

Presentation This work was published in the conference proceedings of "International Conference on Image Analysis and Recognition ICIAR 2019" on August 8, 2019 [125].

6.1 Abstract

This paper proposes a novel deep learning architecture WaveM-CNN for efficient recognition of sub-cellular organelles in microscopic images. Essentially, multi-resolution analysis based on wavelet decomposition and convolution neural network (CNN) are combined in the architecture. In each wavelet transformed sub-space, discriminative features are extracted by convolution kernels to provide various pattern characteristics of the same organelle. The generated feature maps are concatenated and passed directly to the fully connected layers of the classifier. In order to reduce the computational time and improve performance on limited dataset, transfer learning method is adopted, with the utilization of compact MobileNet model. Experiments on two benchmark datasets **CHO** and **2D HeLa** are conducted to evaluate the performance of the proposed model on fluorescence microscopic images of sub-cellular organelles. The classification accuracies of **98.4**% and **96.1**% are achieved on these two datasets respectively, which are significantly higher than both hand-crafted feature based methods and recent deep learning based models.

Keywords: Deep Learning , Multi-resolution, Microscopic Image, Wavelet Transform

6.2 Introduction

One important application of microscopic image analysis is to specify the sub-cellular structures inside a cell where there are proteins of interest. The localization where the proteins are produced by given genes is an essential factor to determine the possible function of such genes. Proteins of similar sequence structures could function differently due to their different compartment localization within the same cell [183].

Microscopic image interpretation is a challenging problem even for experimented pathologists. The output images from typical acquisition systems may have a high resolution but actual objects of interest may only have ten to twenty times smaller resolution [184]. Furthermore, there is usually a strong variability in the organelles shape within the same class, while inter-class variability is relatively small. Thanks to advances in statistical pattern recognition, the acquired microscope images could be automatically and objectively analyzed based on sets of annotated example data.

Algorithms for automatic object classification in microscopic images are essentially composed of feature extraction and classification stages. In conventional hand-crafted feature extraction methods, the set of local characteristics in images such as points, edges or intensity distribution in neighborhood regions are designed specifically for each target dataset [62, 65, 177, 185–188]. Then popular efficient classifiers such as Support Vector Machine (SVM) or Artificial Neural Network (ANN) are applied to determine the class of object based on its extracted features. The classification accuracy by these approaches could be relatively high for microscopic images. However, the remarkable drawback of hand-crafted based methods is time-consuming and difficult to obtain the relevant features for a wide range of datasets: designed parameters are mostly not relevant for new target images other than the specific images for which they were crafted [189]. Recent efforts tried to develop efficient pattern recognition systems inspired by deep learning algorithms. Unlike methods based on hand-crafted feature extraction, deep learning models automatically learn optimal feature representation from image pixel data to perform classification task directly. Noticeable results are recently produced by using CapsNet network [190] or ensembling multiple Convolution Neural Networks [78, 79]. Nevertheless, the reported performance can not overcome the highest result achieved by the previous conventional methods.

This project aims to improve the accuracy compared to recent state of the art deep learning methods and surpass the highest results achieved by conventional hand-crafted feature based methods. Our main contribution in this study are summarized as below:

- Designing a novel multi-resolution architecture, combining 2D wavelet decomposition and the Convolution Neural Network architecture.
- Exploiting compact convolution neural network MobileNet-v1 [191] for automatically extracting features on microscopic image benchmarks.
- Achieving accurate classification of sub-cellular organelles using proposed model with-

out the need for data augmentation.

6.3 Method

This section describes our proposed WaveM-CNN architecture to improve classification accuracy of sub-cellular organelles in microscopic images, as presented in Figure 6.1. First, multi-resolution information is obtained by 2D discrete wavelet transformation. Then, a set of pretrained CNNs is used to extract features in each band of transformed image. Finally, an ensemble network is formed by concatenating generated feature maps to pass through fully connected layers for classification.

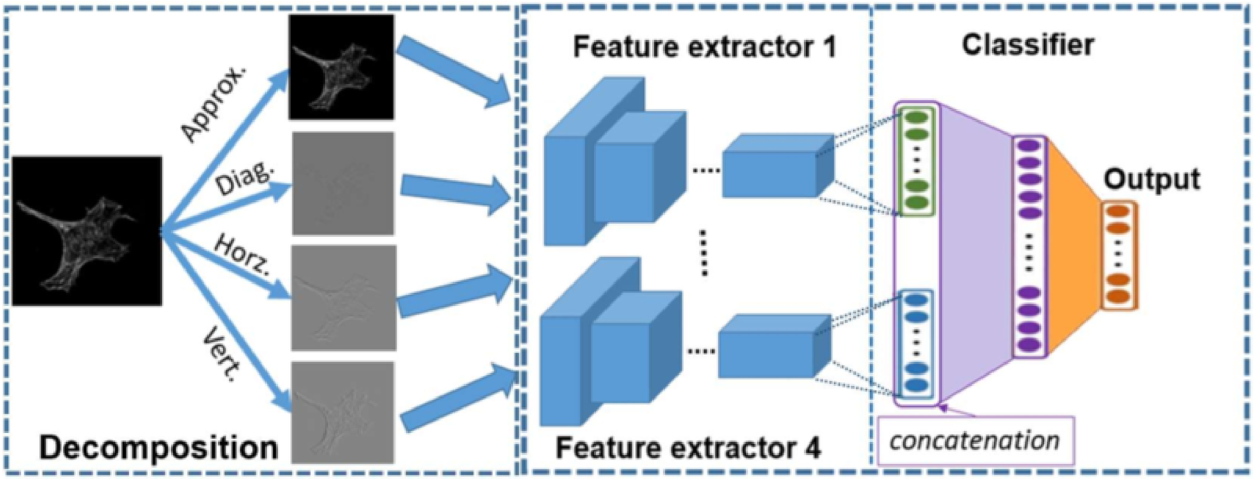


Figure 6.1 The Wavelet based Multi-resolution Convolution Neural Network Architecture

6.3.1 Image decomposition

As the first step, each image is decomposed into a set of filtered images by wavelet transform which is commonly used for representation of an image at multiple levels of resolution or scale. One interesting advantage this wavelet-based multi-resolution analysis is that some features which might not be detected at one resolution could be easily uncovered at another. In this work simple Haar function are applied for the transformation to investigate the effectiveness of proposed approach. Experiments with more sophisticated functions would be done in the future work. The discrete Haar filters are applied on both rows and columns of the image to produce four component images: one approximation and three image details in corresponding horizontal, vertical and diagonal orientations. In general, given the image of size $M \times N = 2^m \times 2^n$, the decomposition outputs at i^th level are calculated by [192]:

$$y_{hh}^{(i)}(u,v) = \sum_{l=1}^{2^{n-i+1}} \left[\sum_{k=1}^{2^{m-i+1}} h(k-2u) y_{hh}^{(i-1)}(k,l) \right] h(l-2v) \qquad (1)$$

$$y_{h\mathrm{g}}^{(i)}(u,v) = \sum_{l=1}^{2^{n-i+1}} \left[\sum_{k=1}^{2^{m-i+1}} h(k-2u) y_{hh}^{(i-1)}(k,l) \right] \mathrm{g}(l-2v) \qquad (2)$$

$$y_{h\mathrm{g}}^{(i)}(u,v) = \sum_{l=1}^{2^{n-i+1}} \left[\sum_{k=1}^{2^{m-i+1}} \mathrm{g}(k-2u) y_{hh}^{(i-1)}(k,l) \right] h(l-2v) \qquad (3)$$

$$y_{\mathrm{gg}}^{(i)}(u,v) = \sum_{l=1}^{2^{n-i+1}} \left[\sum_{k=1}^{2^{m-i+1}} \mathrm{g}(k-2u) y_{hh}^{(i-1)}(k,l) \right] \mathrm{g}(l-2v) \qquad (4)$$

With $u \in \{1, 2, \cdots, 2^{m-i}\}$ and $v \in \{1, 2, \cdots, 2^{n-i}\}$

In the above equations, $y_{hh}^{(i)}$ represents components that contain the approximation coefficients of original image, whereas $y_{hg}^{(i)}$, $y_{gh}^{(i)}$ and $y_{gg}^{(i)}$ are for horizontal details, vertical details and diagonal details coefficients, respectively. $h = \{1/\sqrt{2}, 1/\sqrt{2}\}$ and $g = \{1/\sqrt{2}, -1/\sqrt{2}\}$ represent well known Haar wavelet low-pass filter and high-pass filter. At the initial step: $y_{hh}^{(0)} = \mathit{Original\ image}$ and in this work, only one level of decomposition is performed, thus $i = 1$.

6.3.2 Feature Extraction

After the original images are decomposed into sub-bands, the MobileNet-v1 CNN [191] is used as the feature extractor due to its advantage of small number of parameters and hence, less computation cost. It is a dedicated small architecture introduced by Google for implementation on systems with limited hardware resource, for e.g mobile devices, and thus, it could be integrated on a microscope system. Its principal algorithm is based on *depthwise separable convolution* concept, in which the convolution operation is replaced by *depthwise convolution* followed by *pointwise convolution* [191]. As such, the convolution kernels do not need to have corresponding number of channels to operate on all input channels altogether but just have a single channel to operate on each of them. The generated features maps are then merged by using **1x1** convolution kernels which number depends on the desired output channels. To gain advantage of reduced training time, computation cost and enhanced accuracy, this work utilizes the MobileNet-v1 CNN model which is pretrained on the popular ImageNet dataset [138] of over one million natural images.

6.3.3 Wavelet based multi-resolution CNN architecture

As illustrated in Figure 6.1, a set of four feature extraction modules in the previous section is combined in parallel to carry out the analysis of each decomposed image simultaneously. The feature maps generated independently from all pathways are concatenated to be processed together by the next fully connected layers. This approach of feature combination is found to be more efficient than ensemble methods which use voting scheme with weight averaging or weight learning through training process. Finally, the softmax function is applied at the output to determine the corresponding class of the object. This network model could be trained normally as a single network without dividing the training into multiple steps. Its parameters are initialized following truncated normal distribution with zero mean and standard deviation of 0.001. A learning rate of 0.01 is applied and the optimal number of training epochs is less than 20 epochs with a training batch size of 100 images.

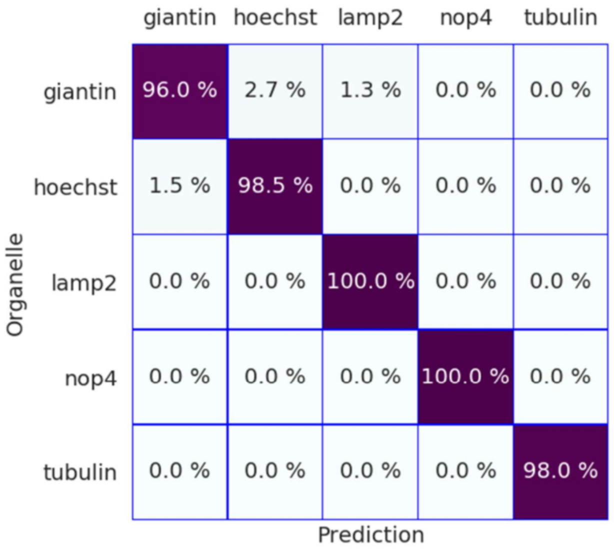
6.3.4 Validation method

The performance of proposed model is evaluated using two benchmark datasets that are publicly available, CHO and 2D HeLa dataset [185]. CHO is a dataset of fluorescence microscopic images of Chinese Hamster Ovary cells, which consists of more than three hundred images produced by five different fluorescent markers. While 2D HeLa contains fluorescence microscopic images of HeLa cells stained with various dyes for targeting specific organelles. There are totally around nine hundreds cell images, with ten different labels. The two benchmark datasets are not suffered from imbalanced classes as there is no significant difference in number of images in each class.

The model performance is validated through conventional five-fold cross-validation method. In each dataset, the whole images are divided by five subsets, each contains 20% of total number of images, collected from all classes. Each time, one subset is used as testing set to measure the accuracy while the remaining subsets are used as training set. In order to avoid serious bias-variance trade-off or over-fitting problem, we follow the *early stopping retraining* policy. First, within the above training set, 75% of the images are used as training data while other 25% of the images are used as validation data. The accuracy performance on validation data is used to determine the optimal number of training epochs for early stopping before over-fitting. Then, the whole training set, including the validation set, is used as the training data to retrain the model, according to the number of training epochs recorded in previous step. The average accuracy after thirty runs, in which image subsets are reshuffled randomly, is the reported classification accuracy to compare with similar previous works.

6.4 Results and Discussion

We first conduct extensive experiments on the CHO dataset according to the validation method described in previous section. The obtained average confusion matrix is shown in Figure 6.2 where the prediction accuracy for any single class is at least 96.0%. In addition, the sensitivity and specificity of the model for each class of CHO dataset are presented in Table 6.1. The lowest sensitivity and specificity among classes are respectively 96.0 % and 99.2%.



	giantin	hoechst	lamp2	nop4	tubulin
giantin	96.0 %	2.7 %	1.3 %	0.0 %	0.0 %
hoechst	1.5 %	98.5 %	0.0 %	0.0 %	0.0 %
lamp2	0.0 %	0.0 %	100.0 %	0.0 %	0.0 %
nop4	0.0 %	0.0 %	0.0 %	100.0 %	0.0 %
tubulin	0.0 %	0.0 %	0.0 %	0.0 %	98.0 %

Figure 6.2 Confusion matrix for CHO classes

In comparison with previous work, as can be seen from Table 6.2, our model classification accuracy is at least 3% higher than the conventional hand-crafted feature extraction based methods. Moreover, its performance outweighs recent deep neural network based models, including ensemble multi-scale CNN network.

A larger dataset of HeLa cell images is subsequently used to validate our model performance. The average confusion matrix for HeLa dataset is shown in Figure 6.3. Obviously, some types of cells are more difficult to distinguish exactly but our model ensures the classification accuracy of more than 90% and most of the time it is over 95%. Table 6.3 represents the sensitivity and specificity values of the model for each class. Noticeably, the specificity is at least 98.9 % for any HeLa organelles.

Early works on classification of HeLa dataset based on hand-crafted feature extraction could achieve relatively good result as indicated in Table 6.4. However, the large number of specific and dedicated feature sets made it difficult to adapt to new set of images, even similar

Table 6.1 Sensitivity and specificity for each CHO class

Class	**giantin**	**hoechst**	**lamp2**	**nop4**	**tubulin**
Sensitivity	96.0 %	98.5 %	100.0 %	100.0 %	98.0 %
Specificity	99.6 %	99.2 %	99.6 %	99.7 %	100.0 %

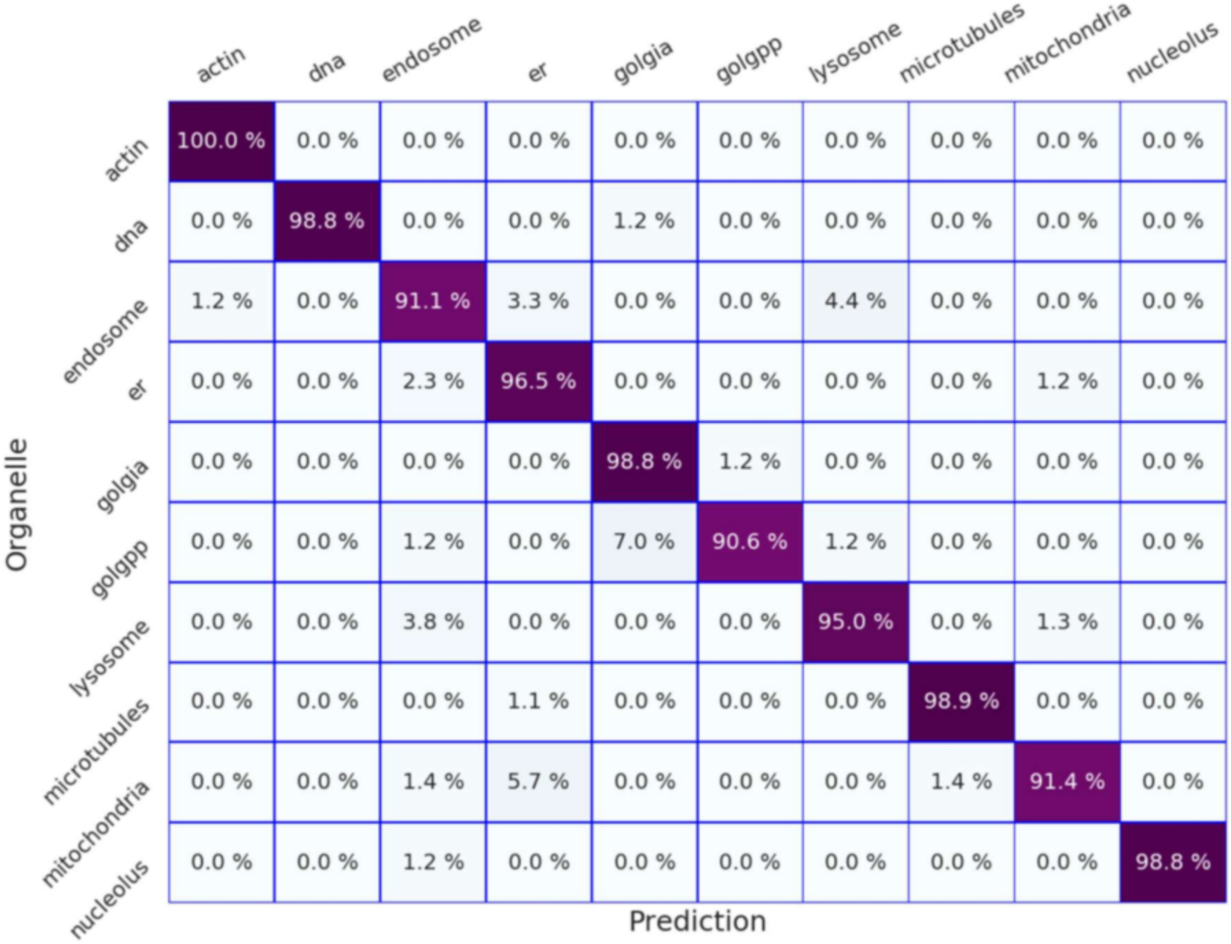
Table 6.2 Performance of various classifiers for CHO images

Type	**Methods**	**Acc.**(%)
Hand-crafted feature extraction based method	Neural network using Zernike moments and Haralick texture features [185]	88.00
	Weighted Neighbor Distances using a Compound Hierarchy of Algorithms Representing Morphology [188]	95.00
Deep learning based method	Single network of AlexNet [78]	29.00
	Single network of GoogleNet [78]	91.00
	Multi-scale convolution neural network with 22 CONV. + 2 FC. [78]	94.00
	Our proposed WaveM-CNN	**98.4**

Table 6.3 Sensitivity and specificity for each HeLa class

Class	**Actin**	**DNA**	**Endosome**	**Golgia**	**Microtubules**
Sensitivity	100.0 %	98.8 %	91.1 %	98.8 %	98.9 %
Specificity	99.9 %	100.0 %	98.9 %	99.1%	99.9 %
	Golgpp	**Lysosome**	**ER**	**Nucleolus**	**Mitochondria**
Sensitivity	90.6 %	95.0 %	96.5 %	98.8 %	91.4 %
Specificity	99.9 %	99.4 %	99.0%	100.0 %	99.7 %

types of organelles. More recent works try to enhance the transfer-ability but they can not avoid degraded performance. In fact, the deficit is very large with regards to the highest benchmark. Table 6.4 also shows recent accuracy levels achieved by Deep learning based methods, including single transfer learning convolution neural network and ensemble network models. For single CNN networks, except for AlexNet, both GoogleNet and Inception-Resnet-v2 could equally produce an accuracy as high as 92%. Although recent CapsNet model achieve the highest accuracy rate of 93.08% but this value is about 3% lower than accuracy rate provided by our WaveM-CNN. Some of recent works also try to apply well known ensemble



Organelle \ Prediction	actin	dna	endosome	er	golgia	golgpp	lysosome	microtubules	mitochondria	nucleolus
actin	100.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
dna	0.0 %	98.8 %	0.0 %	0.0 %	1.2 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
endosome	1.2 %	0.0 %	91.1 %	3.3 %	0.0 %	0.0 %	4.4 %	0.0 %	0.0 %	0.0 %
er	0.0 %	0.0 %	2.3 %	96.5 %	0.0 %	0.0 %	0.0 %	0.0 %	1.2 %	0.0 %
golgia	0.0 %	0.0 %	0.0 %	0.0 %	98.8 %	1.2 %	0.0 %	0.0 %	0.0 %	0.0 %
golgpp	0.0 %	0.0 %	1.2 %	0.0 %	7.0 %	90.6 %	1.2 %	0.0 %	0.0 %	0.0 %
lysosome	0.0 %	0.0 %	3.8 %	0.0 %	0.0 %	0.0 %	95.0 %	0.0 %	1.3 %	0.0 %
microtubules	0.0 %	0.0 %	0.0 %	1.1 %	0.0 %	0.0 %	0.0 %	98.9 %	0.0 %	0.0 %
mitochondria	0.0 %	0.0 %	1.4 %	5.7 %	0.0 %	0.0 %	0.0 %	1.4 %	91.4 %	0.0 %
nucleolus	0.0 %	0.0 %	1.2 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	98.8 %

Figure 6.3 Confusion matrix for HeLa classes

technique on classification task of 2D HeLa images. Two recent works were identified, one uses seven-scale CNN model which is trained from scratch [78] and the other uses triple heterogeneous pretrained CNNs [79]. Even though these methods provide good results, our model achieves around 3.5% gain compared to the best method providing 92.57%.

6.5 Conclusions

We present a novel deep learning architecture that combines the power of multi-resolution analysis by wavelet transform and feature extraction with convolution neural network. The proposed model outperforms previous published works applied on the same datasets of microscopic fluorescent images. Further experiments on other datasets will be conducted to consolidate the transferring ability and generalization of the classification results. This work

Table 6.4 Performance of various classifiers for HeLa images

Type	**Methods**	**Acc.(%)**
Hand-crafted feature extraction based method	Neural Network with a set of 174 features(morphological, Haralick texture, Zernike moments) [65]	91.50
	Neural Network with a setof 26 Haralick texture features [62]	95.30
	SAHLBP (BoW(VQ) + SPM + SVM) [177]	84.49
	SIFT+SAHLBP (BoW(VQ) + SPM + SVM) [177]	86.20
	SIFT(BoW(LLC)+SPM+Softmax) [187]	89.37
Deep learning based method	Single network of AlexNet [78]	11.00
	Single network of GoogleNet [78]	91.00
	Single network of Inception-Resnet-v2 [79]	92.00
	Single network of CapsNet [190]	93.08
	Multi-scale convolution neural network with 22 CONV. + 2 FC. [78]	91.00
	Multiple heterogeneous network of Inception-v3, Resnet152, Inception-Resnet-v2 [79]	92.57
	Our proposed WaveM-CNN	**96.10**

applied the compact network MobileNet-v1, which has much less parameters than well known deep neural networks such as ResNet or Inception.

CHAPTER 7 ARTICLE 4: ORGANET: A ROBUST NETWORK FOR SUBCELLULAR ORGANELLES CLASSIFICATION IN FLUORESCENCE MICROSCOPY IMAGES

Duc Hoa Tran[1], Michel Meunier[2], Farida Cheriet[1]

[1]Department of Computer and Software Engineering, Polytechnique Montréal, Canada
[2]Department of Engineering Physics, Polytechnique Montréal, Canada

Presentation This work was published in the conference proceedings of "42nd International Conference of The IEEE Engineering in Medicine and Biology Society EMBC 2020" on August 27, 2020 [127].

Abstract Automatic identification of subcellular compartments of proteins in fluorescence microscopy images is an important task to quantitatively evaluate cellular processes. A common problem for the development of deep learning based classifiers is that there is only a limited number of labeled images available for training. To address this challenge, we propose a new approach for subcellular organelles classification combining an effective and efficient architecture based on a compact Convolutional Neural Network and deep embedded clustering algorithm. We validate our approach on a benchmark of HeLa cell microscopy images. The network both yields high accuracy that outperforms state of the art methods and has significantly small number of parameters. More interestingly, experimental results show that our method is strongly robust against limited labeled data for training, requiring four times less annotated data than usual while maintaining the high accuracy of 93.9%.

7.1 Introduction

Cells are complex biological structures internally partitioned into compartments which are called organelles. Each organelle contains a specific set of proteins and creates an enclosed environment for their chemical reactions to perform a specific function [193]. Therefore, precise subcellular localization of proteins could provide information about the organelles functions and underlying chemical processes. One of the most suitable approaches for the identification of subcellular organelles is by acquiring fluorescence microscopy images after fluorescent tagging of proteins [65, 193]. However, it is still very challenging in discriminating effectively those organelle types due to their high similarity in image appearance [78]. An automatic and systematic determination of protein locations by a pattern recognition system should enable the possibility of high throughput analysis and minimizes performance inconsistency caused by manual inspection. In previous works, machine learning methods are applied for organelles recognition in florescence microscopy images. In [65], a neural network is used to classify ten major subcellular patterns in HeLa cells, based on the handcrafted extraction of 174 features including morphological features, Haralick textures and Zernike moments. An improved architecture is described in [62] which applied multi-resolution technique and required extraction of 26 Haralick textures from each transformed image. Authors of [194] propose to train a support vector machine (SVM) by using the scale invariant feature transform (SIFT). The main challenge for these handcrafted feature extraction methods is that the feature sets are subjectively designed and often lack of capability to represent complex discriminative structures which results in low classification accuracy. On the other hand, deep learning models based on convolutional neural network (CNN) architecture could automatically learn optimal feature representation directly from raw image data. They have been applied in some recent works, either as a single network [78, 190] or an ensemble of multiple networks [78, 79, 125]. However, they still have limited accuracy and high model complexity in terms of huge number of trainable parameters. Considering the result produced by the conventional handcrafted feature extraction method on the HeLa cell image benchmark [62, 195], there is no published deep convolutional neural network that could outperform the benchmark value [78, 190]. The hybrid model proposed in [125] could achieve better result but a preprocessing step of wavelet transformation is required and the network could not be trained in an end-to-end fashion. In addition, as these deep convolution neural networks have a large number of learnable parameters, the computational cost, memory consumption and the issue of overfitting on limited training data become more serious. As demonstrated in previous works, the supervised deep convolution neural networks require at least 700 labeled images to achieve cross-validation accuracy of around 93%. In this paper, we propose a new efficient deep learning model based on the combination of a lightweight feature extraction structure and deep embedded clustering technique [196] for subcellular organelles recognition in fluorescence microscopy images, denoted as OrgaNet. In contrast to other related works using deep convolution neural network, our design is based on the feature extractor module of Mobilenet-v2 network, which is one of the most compact deep CNN models dedicated for low computational mobile devices [93]. In addition, by using deep embedded clustering, the operation of OrgaNet differs from other works in that output prediction is performed by unsupervised clustering in feature space instead of producing class probability by the softmax function. To better exploit this clustering technique, we formularize an integrated cost function to regularize the feature space during optimization process.

To verify the validity of the proposed method, we design experiments on the benchmark of HeLa cells fluorescence microscopy images. Experimental results demonstrate that this method has achieved high performance which outperforms other state of the art algorithms and maintains stability even with few groundtruth labels.

7.2 Methodology

Fig. 7.1 shows the overall structure of the proposed network. The raw images are fed to an adaptive layer before the feature extractor module to produce compressed feature representation which is then used as input to a fully connected layer where deep embedded clustering is applied for the organelles recognition.

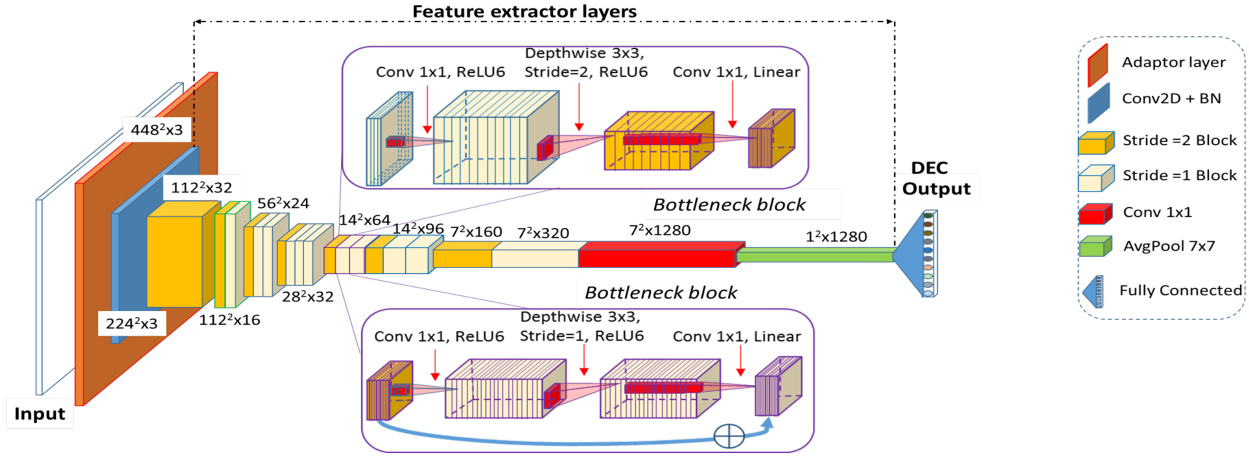


Figure 7.1 The architecture of the proposed method. The numbers on each block represents output dimensions of each layer.

7.2.1 Adaptation layer

The microscopy images are generally larger than spatial input sizes of feature extractor module which is previously designed and trained for natural photos. In particular, each image in 2D HeLa dataset has size of $382 \times 382 \times 1$ whereas the input tensor of the pre-trained feature extractor requires dimension of $224 \times 224 \times 3$. To maintain the high resolution of raw images, instead of down sampling, we add an adaptor layer combining 2D Convolution and Batch Normalization (BN) to normalize activations of this adaptor layer to be zero-mean and of unit standard deviation, which is typically used to help to converge faster and bypass local minima.

7.2.2 Feature extraction

In order to design an efficient deep learning model with low complexity, our network is designed based on the MobileNet-v2 feature extractor architecture [93]. It is developed upon the concept of Depth-wise Separable Convolution, which splits convolution into separate layers: depth-wise convolution and pointwise convolution. Feature representation is captured by using a sequence of bottleneck blocks. Suppose the input tensor to each block has ($height \times width \times channels$) dimensions of ($h \times w \times c$). Each bottleneck block basically consists of three convolutional layers with different characteristics. The first layer uses point-wise (1×1) convolution to expand the number of channels to ($h \times w \times ct$) where expansion factor $t = 6$ which is chosen experimentally. The goal of expansion is to enable the network to represent more complex functions. Then BN and in-place ReLU6 which eliminates all activation values outside the range of $[0, 6]$ to provide a source of non-linearity are used. Then, the second layer performs a lightweight depth-wise convolution to transform lower-level representations, such as pixels, to higher level feature descriptors by applying a single convolutional kernel of size (3×3 per input channel. The stride step parameter could be selected between $s = 1$ or $s = 2$, which respectively maintains or down-samples the input with spatial dimensions ($h \times w$) by half. Similar to the first layer, it also applies BN and ReLU6 function on top of the convolution output. The third layer is again a pointwise (1×1) linear convolution, but without using any non-linear activation transformation, to project high-dimensional extracted feature map back to a low-dimensional representation and is called linear bottleneck. Thus, it embeds the extracted features into a significantly lower-dimensional space and by doing so, the ReLU6 activation function in subsequent bottleneck block will not eliminate too much information while still introducing the needed complexity for the network capability. Whenever the bottleneck block uses down-sampling $s = 2$ in the second convolution layer, a residual connection is used to add the input feature and the linear bottleneck as they are now having the same dimensions. Note that the input feature here is actually the linear bottleneck of the preceding bottleneck block. The motivation of using it is to increase the gradient propagation ability across multiple layers.

7.2.3 Deep embedded clustering (DEC)

Unlike standard classifier networks which produce directly the probability of class label for each input image, we apply method of unsupervised clustering in the latent space to assign class labels. Considering a dataset of n image samples $X = x_1, x_2, ..., x_n$ where each image has $m = height \times width$ pixels or features: $x_i \in R^m$. For each input image x_i, we extract its corresponding latent code y_i which is the output vector at the fully connected layer.

This vector is also the output of the network and it is learnt during the process of network optimization. Then unsupervised clustering is used to group all the $x_i \in X$ into K clusters based on similarity distance calculated on y_i. Compared with xi, the corresponding latent code $yi \in R^d$ has much smaller dimension $d \ll m$, and thus clustering on y_i of reduced dimension is more effective since it reduces "the curse of dimensionality". For example, in our experiment, the microscopy images of HeLa cells have dimension 382×382 but their latent codes have size $d = 10 \times 1$, which allows the K-means clustering algorithm to achieve higher accuracy when working on the latent codes rather than working directly on input images. In this work, after K-means method [197] is used to determine clusters in the latent space, the Hungarian algorithm [198] is applied to map the assigned clusters labels to corresponding groundtruth labels and then produces the classification label for each input image.

7.2.4 Data augmentation

Data augmentation is a common technique to deal with the limited number of available labeled samples for training a deep neural network. To improve generalization ability for target microscopy images of high variability, we attempt to augment our dataset artificially by both geometric and photometric transformations. In particular, for each training image we apply online a random combination of transformations, including rotation, flipping, translating, scaling, shearing, elastic distortion, contrast and brightness perturbation, with a wide range of parameters for each individual transform. Normally, there is a challenge in selecting relevant augmentation types and their parameter ranges to trade-off between capturing the variability of the entire target images and avoiding generating outliers for training which causes data bias problem [199]. However, the negative impact of generated outliers will be significantly mitigated by several mechanisms in our design. Firstly, online augmentation is applied randomly across training iterations instead of offline augmentation which repeatedly stores and feeds the perturbed data at every training iteration. Secondly, momentum based optimization during network training is done via Adaptive Moment Estimation (Adam) which relies on decaying average of past gradients and past squared gradients to compute adaptive learning rates for each network parameter. Finally, the latent space is regulated by normal distribution with zero mean and of unit variance which allows latent variables to converge to separated local clusters.

7.2.5 Network training

In standard training of convolution neural networks, the average cross entropy loss for each training image batch is calculated as:

$$CE = -\frac{1}{B}\sum_{i=1}^{B}\sum_{c=1}^{K} I_{i,c}\log\left(\frac{\exp(y_{i,c})}{\sum_{c=1}^{K}\exp(y_{i,c})}\right) \tag{7.1}$$

Where B is the size of training image mini-batch, K is number of image classes, $I_{i,c}$ is the binary indicator (0 or 1) if true class label c is the correct classification for image sample i, and $y_{i,c}$ is the raw output score produced by the network for an image sample i to have a class label c. In this work, we propose to add an additional constraint for the training which is to match the embedding distribution $q(y)$ of latent variables yi to a prior distribution $p(y)$. In particular, Gaussian distribution with zero mean and unit standard deviation $p(y) \sim \mathcal{N}(0,1)$ is selected as the prior target distribution for optimization. This produces the effect of separating clusters as they are constrained in local spaces along each dimension of the latent variables. The training loss is computed by using embedding trick for Maximum Mean Discrepancy (MMD) [200]:

$$MMD = E_{p(y_i),p(y_j)}[k(y_i, y_j)] + E_{q(y_i),q(y_j)}[k(y_i, y_j)] - 2E_{p(y_i),q(y_j)}[k(y_i, y_j)] \tag{7.2}$$

where

$$k(y_i, y_j) = e^{\frac{||y_i - y_j||^2}{2\sigma^2}}$$

is a kernel to measure the similarity based on Euclidean distance between two samples y_i, y_j which could be drawn from the same or different distributions $p(y), q(y)$. Two distributions are matched if and only if MMD $= 0$. Subsequently, the model is trained to optimize the aggregate cost function by weighted sum of cross-entropy loss and maximum mean discrepancy loss:

$$L = CE + \gamma * MMD \tag{7.3}$$

Where γ is a hyper-parameter to adjust the weight of MMD loss. In our experiments, we set $\gamma = 1$. To train the proposed network, we initially transfer the parameters of the MobileNet-v2 feature extractor which are previously learnt from natural images, while initialize the adaptor layer and final fully connected layer with uniform distribution. Then, Adaptive Moment Estimation (Adam) optimizer is used, with small learning rate of $l_r = 10^{-4}$ which helps stabilize the training loss and avoids destroying entirely the representation of pre-trained feature extractor.

7.3 Experiments and Results

7.3.1 Experiment design

To evaluate the performance of the proposed method, we design experiments with microscopy fluorescence images in 2D HeLa dataset provided in benchmark suite IICBU [195]. There is a total of 862 images of HeLa cells stained with various specific dyes to distinguish 10 different types of intracellular organelles and structures. We investigate the generalization possibility by training our designed network with different amount of available labeled images. In particular, alternatively 20%, 50% and 80% of the total images is used as training data while the rest images are used as testing data in cross-validation measurement. The average performance of 30 runs is reported, with data in each run is randomly reshuffled. For fair comparison among experiments with different training ratios, we also increase the number of training iteration for lower training ratio because the number of augmented data generated online is proportional to the number of training iterations.

7.3.2 Results

Fig. 7.2 shows the performance of our designed network given different amount of training data. Moreover, the impact of data augmentation and deep embedded clustering following MMD regularization is also represented. Our proposed algorithm achieved average classification accuracy of 96.7% when training with 80% of the whole dataset (about 690 images) and 93.9% when training with only 20% of the total images (about 172 images). As the figure shows, data augmentation plays a vital role in training the network, laying the basis for the MMD regularization to further improving performance.

Table 7.1 shows precision and recall values for each organelle class of HeLa cells, corresponding to the case of using 80% dataset for training. Human experts are known to find it extremely difficult to distinguish Endosomes and Lysosomes, and also have trouble to discriminate between Golgpp and Golgi [195]. For these challenges, our proposed method is able to achieve at least 90% precision. When compared to results from other studies, it performs better.

Table 7.2 shows that our proposed model achieves highest classification accuracy when compared with previous best deep learning models in terms of five-fold cross validation. Moreover, it is worth noticing that our model has much less complexity as we did not combine multiple networks together and also the most computation-demanding component is the MobileNet-v2 feature extractor, which is well known for its compactness. The total model size is just $2.24M$ parameters.

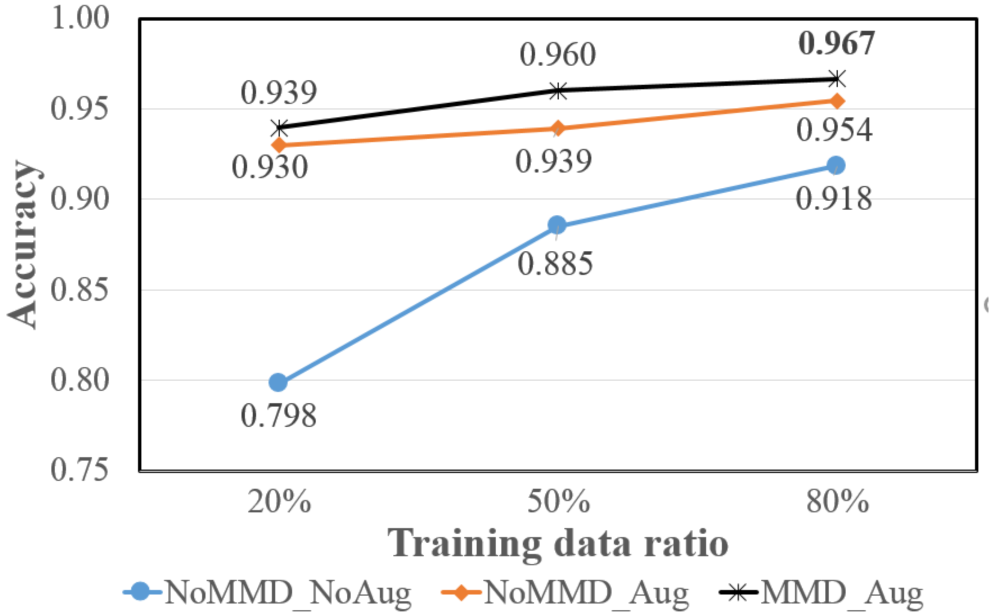


Figure 7.2 Classification performance when training with different amount of labeled data: applying both data augmentatation and MMD regularization for deep embedded clustering (*MMD_Aug*); applying only data augmentation (*NoMMD_Aug*); or not applying any of these two techniques for the network (*NoMMD_NoAug*).

7.4 Conclusion

In this paper, we propose an effective convolution neural network OrgaNet for recognition of subcellular organelles in fluorescent microscopy image. OrgaNet has remarkably small number of parameters and fast computation compared with other published deep learning methods. Our proposed model has surpassed state of the art methods in similar experiment conditions, while it is demonstrated to be very robust when using only a very limited labeled data.

Table 7.1 Precision and recall for all ten classes (%)

Types	**Actin**	**DNA**	**Golgpp**	**ER**	**Golgia**
Precision	100.0	99.7	96.4	96.0	90.9
Recall	99.9	99.9	89.6	97.5	96.7
	Endosome	**Lysosome**	**Microtubules**	**Mitochondria**	**Nucleolus**
Precision	93.6	96.3	99.3	94.4	99.8
Recall	94.9	94.0	98.5	95.5	99.5

Table 7.2 Classification results of different methods

Methods	**Acc.(%)**
Neural Network with 26 Haralick texture features [62]	95.3
GoogleNet [78]	91.0
Inception-Resnet-v2 [79]	92.0
Capsule Network [190]	93.1
Multi-scale CNN network [78]	91.0
Ensemble network of Resnet152, Inception-Resnet-v2, Inception-v3 [79]	92.6
Ensemble of 4 MobileNet-v1 networks and Wavelet Transform preprocessing [125]	96.1
Our proposed OrgaNet	**96.7**

CHAPTER 8 GENERAL DISCUSSION

This thesis project has the general objective of improving the generalization and applicability of deep learning-based models for the analysis of microscopy images. We focus on developing the segmentation and classification algorithms for cells and structures which were captured from various imaging domains and conditions. To overcome the generalization challenge, we develop algorithms that are data-efficient or require the minimum number of labeled data. To address the applicability issue, we propose methods that are computation-efficient and domain-adaptable. The successful developments of these algorithms will assist pathologists or biologists in the diagnostics and analysis of microscopy images in various scenarios. The following sections discuss our findings and methods' limitations.

8.1 Deep learning-based segmentation

Our main contribution is the development of an unsupervised deep learning model for segmentation in histopathological images that is domain-adaptable. Existing models typically offer high performance but are heavily dependent on expensive manual annotation.

In case of the segmentation of nuclei in $H\&E$ stained brightfield images, the annotation of segmentation ground-truth for training deep learning models is not a trivial task. This pixel-level annotation is typically harder than the image-level one in classification task because it could be:

- insufficient: when not all of the objects in an image are labeled, it could provide false negative training signals.
- inconsistent: when the objects within the same category appear differently, they could be annotated differently by a single observer and obviously by different observers.

We have shown that our proposed model eliminates the need for an expensive annotation process while outperforming common unsupervised algorithms and is competitive with some recent supervised methods across diverse image datasets. By self-generating pseudo-data during training process, it also avoids the systematic bias problem in the manual annotation performed by certain observers.

To make the design applicable in a variety of histopathology settings, this work proposes the design approach that is domain-adaptable after deployment. This is very important in practice because the experimental conditions influencing the target data can change frequently, depending on many factors. Meanwhile, the existing domain adaptation and self-supervised methods both assume that target data, even without annotations, are accessible during training. This assumption is often invalid in histopathological settings because images in target domains are generally unknown before training the model and domain shift can occur between different patients [89, 140].

Overall, our proposed segmentation model attained its stated requirements in terms of generalizing well on multiple domains of microscopy images and having the adaptability that eliminates the need for reconfiguration after deployment.

8.1.1 Limitations

Although our proposed solution introduced several contributions to the segmentation topic, the work is still far from complete. The most important shortcoming of our method is the non-competitive instance segmentation scores. After the initial binary segmentation, we use the simple distance transform to obtain the markers for the classical watershed algorithm to produce the final segmentation map. Thus the limitations of the watershed algorithm will hinder directly the overall instance-level segmentation of the framework. A naive approach for mitigating the dependence upon the watershed algorithm and thus increasing instance-level segmentation quality is to have higher resolution input images processed by a deeper encoder-decoder network. In our research prototype, we down-scaled significantly original images that made the touching nuclei phenomenon more challenging. Another potential approach is to use a complementary deep learning network to predict the location of realistic markers [97]. The integration of such an additional network, however, should increase the overall model and timing complexity.

There is also a limitation regarding our modeling algorithm for nuclei morphology. In this study, we considered a simple 2-D Gaussian noise map for the generation of nuclei objects. We assume that the randomly generated shapes represent well the arbitrary variation of nuclei morphology in the image. Although this helps to decrease the bias in the generation of training data, more investigation is needed to better modeling of the distribution of realistic nuclei morphology, such as shapes, area or densities. Accordingly, a more complex algorithm with more parameters to control the quality of the modeling should be developed. One of the alternatives is to reuse available public fluorescent datasets which contain images that are easier for nuclei segmentation [106, 107].

Another drawback is that our model implementation can not achieve real-time performance. As the whole framework contains a deep learning-based training process for each target image, it generally needs to balance the complexity of the network and the running time. If we use a more complex network than the basic U-Net network structure in the model, the segmentation requires more than four minutes.

8.2 Deep learning-based classification

The complexity of a deep learning model is always a principal problem that affects the balance between its ability to learn and to generalize. Since the microscopy image database for each specific application has a limited number of images and ground-truth labels, a computationally complex model will lead to the over-fitting phenomenon. On the other hand, we need to ensure that the model has enough capacity to extract distinguishing feature sets or learn from the data effectively. In addition, we believe that the use of data alone is not sufficient for the existing models to generalize, as also suggested in [84]. In this study, our main contributions are the development of data and computing efficient solutions, based on designing compact network architectures, combining conventional transformation algorithms and regularizing feature space during training from scratch. All of our three proposed classifiers have a significantly smaller number of training parameters but still yield higher accuracy than state-of-the-art methods. Besides facilitating the computation and hardware requirements for training and deploying in real world, our proposed method improved the learning generalization, in cases of weakly supervised learning and multi-domain learning. Thus, the research attained the goal of proposing compact classification algorithms that have strong generalization capability given limited labeled data.

8.2.1 Limitations

Despite several mentioned contributions of our proposed classifiers, there are still certain limitations in our approach.

Firstly, the developed classifiers only work on images of segmented or detected objects. It is often the case that the raw image contains a crowd of cells or organelles, especially in a whole slide image. Thus, it is necessary to have a pre-processing step to separate each of the objects into a single image patch.

Secondly, even though the classifiers can analyze images acquired from different domains or imaging settings, the input images are constrained to fixed predefined dimensions, for example, $224 \times 224 \times 3$. For larger images that are being downscaled to this dimension, their resolution is directly decreased and useful information of the objects may be lost. In addition, it is also not possible to classify high-dimensional images, including multi-spectral or hyperspectral microscopy images.

CHAPTER 9 CONCLUSION

This thesis has addressed the generalization of deep learning models for the analysis, particularly segmentation and classification, of microscopy images. In the proposed approaches, we focus on developing solutions that make the deep learning models generalize well across multiple domains of application and imaging conditions. Firstly, we proposed a novel unsupervised learning model for the segmentation of nuclei in $H\&E$ stained images. In addition to the elimination of manual annotation, it provides the advantage of domain-adaptability. Therefore, the framework is practically applicable for a wide range of histopathology settings. Another contribution to this study is the development of an integrated optimization function to improve network adaptability to the target image. We believe the developed method will contribute to increasing analysis efficiency and application range, and trigger other projects to remove the dependence on manual annotation. As a potential expansion, the use of direct feedback by pathologists could be investigated to integrate into the model design. It is also advisable to assess how the continual learning within a targeted laboratory could improve the performance, i.e. by fine-tuning a network previously trained on patients' samples acquired in a laboratory setting. These developments will improve the reliability of the model and the confidence of pathologists when using our automatic segmentation algorithm.

Secondly, we propose several methods for the classification of various types of objects in microscopy images. We started with the development of a lightweight CNN classifier, entitled Mobincep, for learning multiple domains of microscopy images. It could work effectively on images that capture different levels of cellular structures, without requiring the adaptation of domain-specific parameters. Moreover, we also formulate a simple yet effective optimization function and devise a suitable training pipeline allowing our network to outperform state-of-the-art methods either on each separate domain or on all domains at once. Our proposed model is remarkably compact and robust against limited available training data. Because of its low complexity and being widely applicable, the approach becomes more appealing for deployment in clinical and biomedical studies.

Instead of the above generic model for multi-domain images, we then investigated the scenario where we want to focus solely on improving further the accuracy for the recognition of subcellular organelles. We first described a novel deep learning architecture WaveM-CNN that combines the power of conventional multi-resolution analysis and automatic feature extraction from a deep learning network. The proposed approach significantly improves the classification performance compared with state-of-the-art deep learning models on the same datasets of microscopic fluorescent images. Unlike existing methods, we developed our method on top of a lightweight neural network, which has dramatically fewer parameters than well-known deep neural networks.

Following this model, we extended the research with the development of an effective convolution neural network OrgaNet which also has a remarkably small number of parameters and fast computation compared with other published deep learning methods. Unlike the previous WaveM-CNN model in which the feature extractor layers were pre-trained on the popular ImageNet dataset of natural photos, we decided to train Organet from scratch, without the use of any external images. This approach is particularly useful when there is no available pre-trained model or when the pre-trained feature extractor is not effective for the target microscopy images. In particular, we formulated a new optimization function to regulate the feature space and devised a suitable optimization procedure. We demonstrate that our proposed model has surpassed state-of-the-art methods in similar experimental conditions. More importantly, the model can generalize well as it requires very limited labeled data, about four times less than other approaches, to achieve a similar level of accuracy.

We believe the developed methods will contribute to significantly decreasing the burden of generating ground-truth annotation for training data and increase the analysis efficiency of pathologists and biomedical scientists. The limitations mentioned in chapter 8 create opportunities for new developments. It would be very helpful to have a complete automatic pipeline that can first perform segmentation or detection and then classify the objects of interest. Some existing studies have proposed such an end-to-end framework. However, their studies just focus on a specific application with limited transferability to others. A multi-domain framework is practically needed to save the time and labor of pathologists and biology scientists. An end-to-end framework also mitigates the requirement of normalizing the size of input images for the classifiers. For example, the output dimensions of detected objects images are constrained to fixed values, unlike the case where the detection bounding box is done manually by experts.

Another interesting research topic could be the development of a continual learning method. This helps a deep learning agent to leverage the knowledge accumulated from learning previous tasks and thus learn well a new task given little labeled data [201]. In principle, continual learning aims at two properties: (1) avoiding forgetting or the degradation of performance on a task learned in the past and (2) better learning over time. When developing our multi-domain classifiers, we implied the static nature of the learning problem, assuming that pathologists or biomedical scientists have only three certain imaging domains. In practice, however, new tasks may arrive and the deployed agent can not be used. In another scenario, pathologists may want to use the agent for the classification of new types of cancer. The transfer learning could be applied, but the agent may fail to classify previous types of samples.

REFERENCES

[1] OpenStax *et al.*, “Microbiology,” Available: https://openstax.org/details/books/microbiology, 2016.

[2] L. A. Urry, *Campbell biology / Lisa A. Urry, Michael L. Cain, Steven A. Wasserman, Peter V. Minorsky, Jane B. Reece.*, eleventh edition. ed. New York, NY: Pearson Education, Inc., 2017 - 2017.

[3] R. C. Gonzalez and R. E. Woods, *Digital Image Processing (3rd Edition)*. USA: Prentice-Hall, Inc., 2006.

[4] F. Xing *et al.*, “Deep learning in microscopy image analysis: A survey,” *IEEE Trans. Neural Networks Learn. Syst.*, vol. 29, no. 10, pp. 4550–4568, 2018.

[5] Y. Lecun *et al.*, “Gradient-based learning applied to document recognition,” in *Proceedings of the IEEE*, 1998, pp. 2278–2324.

[6] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, ser. Lecture Notes in Computer Science, N. Navab *et al.*, Eds., vol. 9351. Springer, 2015, pp. 234–241. [Online]. Available: https://doi.org/10.1007/978-3-319-24574-4_28

[7] J. Heaton, “Ian goodfellow, yoshua bengio, and aaron courville: Deep learning - the MIT press, 2016, 800 pp, ISBN: 0262035618,” *Genet. Program. Evolvable Mach.*, vol. 19, no. 1-2, pp. 305–307, 2018.

[8] M. F. Stollenga *et al.*, “Parallel multi-dimensional lstm, with application to fast biomedical volumetric image segmentation,” in *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, C. Cortes *et al.*, Eds., 2015, pp. 2998–3006. [Online]. Available: https://proceedings.neurips.cc/paper/2015/hash/d43ab110ab2489d6b9b2caa394bf920f-Abstract.html

[9] X. Li *et al.*, “Multi-layer domain adaptation method for rolling bearing fault diagnosis,” *Signal Process.*, vol. 157, pp. 180–197, 2019. [Online]. Available: https://doi.org/10.1016/j.sigpro.2018.12.005

[10] E. Meijering, “Cell segmentation: 50 years down the road [life sciences],” *IEEE Signal Process. Mag.*, vol. 29, no. 5, pp. 140–145, 2012. [Online]. Available: https://doi.org/10.1109/MSP.2012.2204190

[11] Y. al kofahi *et al.*, “A deep learning-based algorithm for 2-d cell segmentation in microscopy images,” *BMC Bioinformatics*, vol. 19, 10 2018.

[12] Z. Liu *et al.*, “A survey on applications of deep learning in microscopy image analysis,” *Computers in biology and medicine*, vol. 134, p. 104523, 2021.

[13] Q. Wu, F. A. Merchant, and K. R. Castleman, “Microscope image processing,” 2010.

[14] A. Mescher *et al.*, *Junqueiraś Basic Histology, 15th edition, 2018*, 01 2018.

[15] M. Lafarge *et al.*, “Learning domain-invariant representations of histological images,” *Frontiers in Medicine*, vol. 6, 07 2019.

[16] Y. Rivenson *et al.*, “Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning,” *Nature Biomedical Engineering*, vol. 3, pp. 466–477, 2019.

[17] J. Rony *et al.*, “Deep weakly-supervised learning methods for classification and localization in histology images: a survey,” *CoRR*, vol. abs/1909.03354, 2019. [Online]. Available: http://arxiv.org/abs/1909.03354

[18] A. X. Lu *et al.*, “Yeastspotter: accurate and parameter-free web segmentation for microscopy images of yeast cells,” *Bioinform.*, vol. 35, no. 21, pp. 4525–4527, 2019. [Online]. Available: https://doi.org/10.1093/bioinformatics/btz402

[19] J. Caicedo *et al.*, “Nucleus segmentation across imaging experiments: the 2018 data science bowl,” *Nature Methods*, vol. 16, 12 2019.

[20] V. Ljosa, K. Sokolnicki, and A. Carpenter, “Annotated high-throughput microscopy image sets for validation,” *Nature methods*, vol. 9, p. 637, 06 2012.

[21] G. Jacquemet, “Deep learning to analyse microscopy images,” *The Biochemist*, vol. 43, no. 5, pp. 60–64, 08 2021.

[22] G. Litjens *et al.*, “A survey on deep learning in medical image analysis,” *Medical Image Anal.*, vol. 42, pp. 60–88, 2017. [Online]. Available: https://doi.org/10.1016/j.media.2017.07.005

[23] L. von Chamier *et al.*, “Democratising deep learning for microscopy with zerocostdl4mic,” *Nature Communications*, vol. 12, 2021.

[24] N. Efford, *Digital Image Processing: A Practical Introduction Using Java (with CD-ROM)*, 1st ed. USA: Addison-Wesley Longman Publishing Co., Inc., 2000.

[25] Y. Al-Kofahi *et al.*, "Improved automatic detection and segmentation of cell nuclei in histopathology images," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 4, pp. 841–852, 2010. [Online]. Available: https://doi.org/10.1109/TBME.2009.2035102

[26] O. Schmitt and M. H. Hasse, "Morphological multiscale decomposition of connected regions with emphasis on cell clusters," *Comput. Vis. Image Underst.*, vol. 113, pp. 188–201, 2009.

[27] P. Delmas, "Compsci 773 - intelligent vision systems course," https://www.cs.auckland.ac.nz/courses/compsci773s1c/, 2018.

[28] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[29] F. Xing and L. Yang, "Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: A comprehensive review," *IEEE Reviews in Biomedical Engineering*, vol. 9, pp. 234–263, 2016.

[30] F. Meyer and S. Beucher, "Morphological segmentation," *J. Vis. Commun. Image Represent.*, vol. 1, no. 1, pp. 21–46, 1990. [Online]. Available: https://doi.org/10.1016/1047-3203(90)90014-M

[31] "Active contour model," Available: https://en.wikipedia.org/wiki/Active_contour_model, accessed: 2022-02-02.

[32] H. Cai *et al.*, "Repulsive force based snake model to segment and track neuronal axons in 3d microscopy image stacks," *NeuroImage*, vol. 32, no. 4, pp. 1608–1620, 2006. [Online]. Available: https://doi.org/10.1016/j.neuroimage.2006.05.036

[33] F. Xing and L. Yang, "Robust selection-based sparse shape model for lung cancer image segmentation," *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 16 Pt 3, pp. 404–12, 2013.

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*,

P. L. Bartlett *et al.*, Eds., 2012, pp. 1106–1114. [Online]. Available: https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

[35] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[36] Y. LeCun, Y. Bengio, and G. E. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. [Online]. Available: https://doi.org/10.1038/nature14539

[37] C. Szegedy *et al.*, "Going deeper with convolutions," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.

[38] K. He *et al.*, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016.* IEEE Computer Society, 2016, pp. 770–778. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90

[39] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2010. [Online]. Available: https://doi.org/10.1109/TKDE.2009.191

[40] I. J. Goodfellow, Y. Bengio, and A. C. Courville, *Deep Learning*, ser. Adaptive computation and machine learning. MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org/

[41] F. Ning *et al.*, "Toward automatic phenotyping of developing embryos from videos," *IEEE Trans. Image Process.*, vol. 14, no. 9, pp. 1360–1371, 2005. [Online]. Available: https://doi.org/10.1109/TIP.2005.852470

[42] D. C. Ciresan *et al.*, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, P. L. Bartlett *et al.*, Eds., 2012, pp. 2852–2860. [Online]. Available: https://proceedings.neurips.cc/paper/2012/hash/459a4ddcb586f24efd9395aa7662bc7c-Abstract.html

[43] Y. Song *et al.*, "Accurate segmentation of cervical cytoplasm and nuclei based on multiscale convolutional network and graph partitioning," *IEEE Transactions on Biomedical Engineering*, vol. 62, pp. 2421–2433, 2015.

[44] H. Su *et al.*, “Region segmentation in histopathological breast cancer images using deep convolutional neural network,” *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pp. 55–58, 2015.

[45] M. Sapkota *et al.*, “Automatic muscle perimysium annotation using deep convolutional neural network,” *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pp. 205–208, 2015.

[46] A. Fakhry, H. Peng, and S. Ji, “Deep models for brain EM image segmentation: novel insights and improved performance,” *Bioinform.*, vol. 32, no. 15, pp. 2352–2358, 2016. [Online]. Available: https://doi.org/10.1093/bioinformatics/btw165

[47] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015.* IEEE Computer Society, 2015, pp. 3431–3440. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7298965

[48] H. Chen *et al.*, “Deep contextual networks for neuronal structure segmentation,” in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, D. Schuurmans and M. P. Wellman, Eds. AAAI Press, 2016, pp. 1167–1173. [Online]. Available: http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11789

[49] S. Ren *et al.*, “Faster R-CNN: towards real-time object detection with region proposal networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017. [Online]. Available: https://doi.org/10.1109/TPAMI.2016.2577031

[50] K. He *et al.*, “Mask R-CNN,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 386–397, 2020. [Online]. Available: https://doi.org/10.1109/TPAMI.2018.2844175

[51] J. W. Johnson, “Adapting mask-rcnn for automatic nucleus segmentation,” *CoRR*, vol. abs/1805.00500, 2018. [Online]. Available: http://arxiv.org/abs/1805.00500

[52] S. Fujita and X. Han, “Cell detection and segmentation in microscopy images with improved mask R-CNN,” in *Computer Vision - ACCV 2020 Workshops - 15th Asian Conference on Computer Vision, Kyoto, Japan, November 30 - December 4, 2020, Revised Selected Papers*, ser. Lecture Notes in Computer Science, I. Sato and B. Han, Eds., vol. 12628. Springer, 2020, pp. 58–70. [Online]. Available: https://doi.org/10.1007/978-3-030-69756-3_5

[53] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[54] K. Cho *et al.*, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, A. Moschitti, B. Pang, and W. Daelemans, Eds. ACL, 2014, pp. 1724–1734. [Online]. Available: https://doi.org/10.3115/v1/d14-1179

[55] Y. Xie *et al.*, “Spatial clockwork recurrent neural network for muscle perimysium segmentation,” in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016 - 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II*, ser. Lecture Notes in Computer Science, S. Ourselin *et al.*, Eds., vol. 9901, 2016, pp. 185–193. [Online]. Available: https://doi.org/10.1007/978-3-319-46723-8_22

[56] A. Courville, “Ift6135- representation learning course,” https://sites.google.com/mila.quebec/ift6135, 2019.

[57] P. Vincent *et al.*, “Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion,” *J. Mach. Learn. Res.*, vol. 11, pp. 3371–3408, 2010.

[58] H. Su *et al.*, “Robust cell detection and segmentation in histopathological images using sparse reconstruction and stacked denoising autoencoders,” *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 9351, pp. 383–390, 2017.

[59] F. Mahmood *et al.*, “Deep adversarial training for multi-organ nuclei segmentation in histopathology images,” *IEEE Trans. Medical Imaging*, vol. 39, no. 11, pp. 3257–3267, 2020. [Online]. Available: https://doi.org/10.1109/TMI.2019.2927182

[60] T. Ojala, M. Pietikäinen, and T. Mäenpää, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, pp. 971–987, 2002.

[61] N. A. Hamilton *et al.*, “Fast automated cell phenotype image classification,” *BMC Bioinform.*, vol. 8, 2007. [Online]. Available: https://doi.org/10.1186/1471-2105-8-110

[62] A. Chebira et al., “A multiresolution approach to automated classification of protein subcellular location images,” *BMC Bioinformatics*, vol. 8, no. 1, p. 210, Jun 2007.

[63] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995. [Online]. Available: https://doi.org/10.1007/BF00994018

[64] M. H. Hassoun, *Fundamentals of Artificial Neural Networks*, 1st ed. Cambridge, MA, USA: MIT Press, 1995.

[65] K. Huang et al., “Boosting accuracy of automated classification of fluorescence microscope images for location proteomics,” *BMC Bioinformatics*, vol. 5, no. 1, p. 78, Jun 2004.

[66] T. Pärnamaa and L. Parts, “Accurate classification of protein subcellular localization from high throughput microscopy images using deep learning,” *G3 (Bethesda, Md.)*, vol. 7, 04 2017.

[67] J. Jin *et al.*, “Deep learning of diffraction image patterns for accurate classification of 5 cell types.” *Journal of biophotonics*, p. e2389, 2019.

[68] H. Niioka *et al.*, “Classification of c2c12 cells at differentiation by convolutional neural network of deep learning using phase contrast images,” *Human Cell*, vol. 31, pp. 87–93, 2017.

[69] A. Witmer and B. Bhanu, “Multi-label classification of stem cell microscopy images using deep learning,” *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 1408–1413, 2018.

[70] C. L. Chen *et al.*, “Deep learning in label-free cell classification,” *Scientific Reports*, vol. 6, 2016.

[71] F. wei Qin *et al.*, “Fine-grained leukocyte classification with deep residual learning for microscopic images,” *Computer methods and programs in biomedicine*, vol. 162, pp. 243–252, 2018.

[72] A. I. Shahin *et al.*, “White blood cells identification system based on convolutional deep neural learning networks,” *Computer methods and programs in biomedicine*, vol. 168, pp. 69–80, 2019.

[73] L. Alzubaidi *et al.*, “Deep learning models for classification of red blood cells in microscopy images to aid in sickle cell anemia diagnosis,” *Electronics*, vol. 9, p. 427, 2020.

[74] N. Coudray *et al.*, "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning," *Nature Medicine*, vol. 24, pp. 1559–1567, 2018.

[75] S. U. Khan *et al.*, "A novel deep learning based framework for the detection and classification of breast cancer using transfer learning," *Pattern Recognit. Lett.*, vol. 125, pp. 1–6, 2019.

[76] M. S. Iqbal *et al.*, "Efficient cell classification of mitochondrial images by using deep learning," *Journal of Optics*, vol. 48, pp. 113–122, 2019.

[77] Z. Gao *et al.*, "Hep-2 cell image classification with deep convolutional neural networks," *IEEE J. Biomed. Health Informatics*, vol. 21, no. 2, pp. 416–428, 2017.

[78] W. J. Godinez *et al.*, "A multi-scale convolutional neural network for phenotyping high-content cellular images," *Bioinformatics*, vol. 33, p. 2010–2019, 2017.

[79] L. Nguyen *et al.*, "Deep cnns for microscopic image classification by exploiting transfer learning and feature concatenation," *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, 2018.

[80] A. A. Cruz-Roa *et al.*, "A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2013 - 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part II*, ser. Lecture Notes in Computer Science, K. Mori *et al.*, Eds., vol. 8150. Springer, 2013, pp. 403–410. [Online]. Available: https://doi.org/10.1007/978-3-642-40763-5_50

[81] H. Chang *et al.*, "Stacked predictive sparse coding for classification of distinct regions in tumor histopathology," in *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013.* IEEE Computer Society, 2013, pp. 169–176. [Online]. Available: https://doi.org/10.1109/ICCV.2013.28

[82] D. de Ridder, J. de Ridder, and M. J. T. Reinders, "Pattern recognition in bioinformatics," *Briefings in bioinformatics*, vol. 14 5, pp. 633–47, 2013.

[83] P. Mehta *et al.*, "A high-bias, low-variance introduction to machine learning for physicists," *CoRR*, vol. abs/1803.08823, 2018. [Online]. Available: http://arxiv.org/abs/1803.08823

[84] V. S. Zinchuk and O. Grossenbacher-Zinchuk, "Machine learning for analysis of microscopy images: A practical guide," *Current Protocols in Cell Biology*, vol. 86, 2020.

[85] K. Kawaguchi, L. P. Kaelbling, and Y. Bengio, "Generalization in deep learning," *CoRR*, vol. abs/1710.05468, 2017. [Online]. Available: http://arxiv.org/abs/1710.05468

[86] C. Zhang *et al.*, "Understanding deep learning requires rethinking generalization," *ArXiv*, vol. abs/1611.03530, 2017.

[87] A. NG, "Cs229-machine learning course," https://cs229.stanford.edu/, 2019.

[88] R. Grosse, "Csc321-neural networks and machine learning - lecture 9 - generalization," https://www.cs.toronto.edu/~lczhang/321, 2020.

[89] K. Zhou *et al.*, "Domain generalization: A survey," *CoRR*, vol. abs/2103.02503, 2021. [Online]. Available: https://arxiv.org/abs/2103.02503

[90] J.-Y. Zhu *et al.*, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251.

[91] L. Alzubaidi *et al.*, "Review of deep learning: concepts, cnn architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, 2021.

[92] A. G. Howard *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *CoRR*, vol. abs/1704.04861, 2017.

[93] M. Sandler *et al.*, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *2018 IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[94] N. Srivastava *et al.*, "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014. [Online]. Available: http://dl.acm.org/citation.cfm?id=2670313

[95] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, F. R. Bach and D. M. Blei, Eds., vol. 37. JMLR.org, 2015, pp. 448–456.

[96] "The cancer genome atlas," Available: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga, accessed: 2021-12-22.

[97] N. Kumar *et al.*, “A multi-organ nucleus segmentation challenge,” *IEEE Transactions on Medical Imaging*, vol. 39, no. 5, pp. 1380–1391, 2020.

[98] M. Majurski *et al.*, “Impact of sampling and augmentation on generalization accuracy of microscopy image segmentation methods.” Computer Vision for Microscopy Image Analysis (CVMI), Salt Lake City, UT, 2018-06-22 2018. [Online]. Available: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=925809

[99] K. K. Jha and H. Sekhar Dutta, “Mutual information based hybrid model and deep learning for acute lymphocytic leukaemia detection in single cell blood smear images,” *Computer Methods and Programs in Biomedicine*, vol. 179, p. 104987, 07 2019.

[100] C. Xiao *et al.*, “Automatic mitochondria segmentation for em data using a 3d supervised convolutional network,” *Frontiers in Neuroanatomy*, vol. 12, 11 2018.

[101] M. Žerovnik Mekuč *et al.*, “Automatic segmentation of mitochondria and endolysosomes in volumetric electron microscopy data,” *Computers in Biology and Medicine*, 03 2020.

[102] Y. Kassim *et al.*, “Deep learning segmentation for epifluorescence microscopy images,” *Microscopy and Microanalysis*, vol. 23, pp. 140–141, 07 2017.

[103] I. Ud Din, N. Islam, and J. Rodrigues, “A novel deep learning based framework for the detection and classification of breast cancer using transfer learning,” *Pattern Recognition Letters*, vol. 125, pp. 1–6, 06 2019.

[104] K. Yao, N. Rochman, and S. Sun, “Cell type classification and unsupervised morphological phenotyping from low-resolution images using deep learning,” *Scientific Reports*, vol. 9, pp. 1–13, 09 2019.

[105] S. Kecheril Sadanandan *et al.*, “Automated training of deep convolutional neural networks for cell segmentation,” *Scientific Reports*, vol. 7, 12 2017.

[106] R. Hollandi *et al.*, “nucleaizer: A parameter-free deep learning framework for nucleus segmentation using image style transfer,” *Cell Systems*, vol. 10, 05 2020.

[107] D. Liu *et al.*, “Unsupervised instance segmentation in microscopy images via panoptic domain adaptation and task re-weighting,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020.* Computer Vision Foundation / IEEE, 2020, pp. 4242–4251.

[108] M. Sahasrabudhe *et al.*, “Self-supervised nuclei segmentation in histopathological images using attention,” in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part V*, ser. Lecture Notes in Computer Science, A. L. Martel *et al.*, Eds., vol. 12265. Springer, 2020, pp. 393–402. [Online]. Available: https://doi.org/10.1007/978-3-030-59722-1_38

[109] Y. Huang *et al.*, “Epithelium-stroma classification via convolutional neural networks and unsupervised domain adaptation in histopathological images,” *IEEE J. Biomed. Health Informatics*, vol. 21, no. 6, pp. 1625–1632, 2017.

[110] J. Ren *et al.*, “Adversarial domain adaptation for classification of prostate histopathology whole-slide images,” in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II*, ser. Lecture Notes in Computer Science, A. F. Frangi *et al.*, Eds., vol. 11071. Springer, 2018, pp. 201–209. [Online]. Available: https://doi.org/10.1007/978-3-030-00934-2_23

[111] L. Hou *et al.*, “Robust histopathology image analysis: To label or to synthesize?” in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019.* Computer Vision Foundation / IEEE, 2019, pp. 8533–8542. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Hou_Robust_Histopathology_Image_Analysis_To_Label_or_to_Synthesize_CVPR_2019_paper.html

[112] K. He *et al.*, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016.* IEEE Computer Society, 2016, pp. 770–778.

[113] ——, “Mask r-cnn,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[114] S. e. A. Raza *et al.*, “Micro-net: A unified model for segmentation of various objects in microscopy images,” *Medical Image Analysis*, vol. 52, 04 2018.

[115] S. Graham and N. M. Rajpoot, “Sams-net: Stain-aware multi-scale network for instance-based nuclei segmentation in histology images,” in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 2018, pp. 590–594.

[116] H. Chen *et al.*, “Dcan: Deep contour-aware networks for accurate gland segmentation,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 06 2016, pp. 2487–2496.

[117] Y. Cui *et al.*, “A deep learning algorithm for one-step contour aware nuclei segmentation of histopathological images,” *Medical & Biological Engineering & Computing*, vol. 57, 03 2018.

[118] S. Graham *et al.*, “Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images,” *Medical Image Analysis*, vol. 58, p. 101563, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841519301045

[119] A. Lagree *et al.*, “A review and comparison of breast tumor cell nuclei segmentation performances using deep convolutional neural networks,” *Scientific Reports*, vol. 11, p. 8025, 04 2021.

[120] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *Proceedings of ICLR*, pp. 1–14, 2015.

[121] G. Huang *et al.*, “Densely connected convolutional networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269.

[122] K. He *et al.*, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[123] S. Min, B. Lee, and S. Yoon, “Deep learning in bioinformatics,” *Briefings Bioinform.*, vol. 18, no. 5, pp. 851–869, 2017. [Online]. Available: https://doi.org/10.1093/bib/bbw068

[124] C. Szegedy *et al.*, “Inception-v4, inception-resnet and the impact of residual connections on learning,” in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, S. Singh and S. Markovitch, Eds. AAAI Press, 2017, pp. 4278–4284. [Online]. Available: http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14806

[125] D. H. Tran, M. Meunier, and F. Cheriet, “Wavem-cnn for automatic recognition of sub-cellular organelles,” in *Image Analysis and Recognition - 16th International Conference, ICIAR 2019, Waterloo, ON, Canada, August 27-29, 2019, Proceedings, Part I*, ser. Lecture Notes in Computer Science, F. Karray, A. Campilho, and

A. C. H. Yu, Eds., vol. 11662. Springer, 2019, pp. 186–194. [Online]. Available: https://doi.org/10.1007/978-3-030-27202-9_16

[126] M. Ghifary *et al.*, "Scatter component analysis: A unified framework for domain adaptation and domain generalization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 7, pp. 1414–1430, 2017. [Online]. Available: https://doi.org/10.1109/TPAMI.2016.2599532

[127] D. H. Tran, M. Meunier, and F. Cheriet, "Organet: A robust network for subcellular organelles classification in fluorescence microscopy images," in *42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC 2020, Montreal, QC, Canada, July 20-24, 2020.* IEEE, 2020, pp. 1887–1890. [Online]. Available: https://doi.org/10.1109/EMBC44109.2020.9175162

[128] ——, "Deep image clustering using self-learning optimization in a variational auto-encoder," in *Pattern Recognition. ICPR International Workshops and Challenges - Virtual Event, January 10-15, 2021, Proceedings, Part II*, ser. Lecture Notes in Computer Science, A. D. Bimbo *et al.*, Eds., vol. 12662. Springer, 2020, pp. 736–749. [Online]. Available: https://doi.org/10.1007/978-3-030-68790-8_56

[129] D. N. Louis *et al.*, "Computational Pathology: A Path Ahead," *Archives of Pathology and Laboratory Medicine*, vol. 140, no. 1, pp. 41–50, 06 2015. [Online]. Available: https://doi.org/10.5858/arpa.2015-0093-SA

[130] N. Kumar *et al.*, "A dataset and a technique for generalized nuclear segmentation for computational pathology," *IEEE Trans. Medical Imaging*, vol. 36, no. 7, pp. 1550–1560, 2017. [Online]. Available: https://doi.org/10.1109/TMI.2017.2677499

[131] A. Mescher *et al.*, *Junqueira's Basic Histology, 15th edition, 2018*, 01 2018.

[132] H. Qu *et al.*, "Weakly supervised deep nuclei segmentation using partial points annotation in histopathology images," *IEEE Transactions on Medical Imaging*, vol. 39, no. 11, pp. 3655–3666, 2020.

[133] H. Jung, B. Lodhi, and J. Kang, "An automatic nuclei segmentation method based on deep convolutional neural networks for histopathology images," *BMC Biomedical Engineering*, vol. 1, no. 1, p. 24, Oct 2019. [Online]. Available: https://doi.org/10.1186/s42490-019-0026-8

[134] H. Irshad *et al.*, "Methods for nuclei detection, segmentation, and classification in digital histopathology: A review—current status and future potential," *IEEE Reviews in Biomedical Engineering*, vol. 7, pp. 97–114, 2014.

[135] F. Yi *et al.*, "Automatic extraction of cell nuclei from H&E-stained histopathological images," *Journal of Medical Imaging*, vol. 4, no. 2, pp. 1 – 12, 2017. [Online]. Available: https://doi.org/10.1117/1.JMI.4.2.027502

[136] Y. Al-Kofahi *et al.*, "Improved automatic detection and segmentation of cell nuclei in histopathology images," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 841–852, 2010.

[137] F. Xing *et al.*, "Deep learning in microscopy image analysis: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 10, pp. 4550–4568, 2018.

[138] J. Deng *et al.*, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[139] R. F. Laine *et al.*, "Avoiding a replication crisis in deep-learning-based bioimage analysis," *Nature Methods*, vol. 18, no. 10, pp. 1136–1144, Oct 2021. [Online]. Available: https://doi.org/10.1038/s41592-021-01284-3

[140] J. Wang *et al.*, "Generalizing to unseen domains: A survey on domain generalization," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, Z. Zhou, Ed. ijcai.org, 2021, pp. 4627–4635. [Online]. Available: https://doi.org/10.24963/ijcai.2021/628

[141] L. Hou *et al.*, "Robust histopathology image analysis: To label or to synthesize?" in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8525–8534.

[142] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[143] S. Beucher and C. Lantuéjoul, "Use of watersheds in contour detection," vol. 132, 01 1979.

[144] A. Ruifrok and D. Johnston, "Quantification of histochemical staining by color deconvolution," *Anal Quant Cytol Histol*, vol. 23, 01 2001.

[145] G. French, A. Oliver, and T. Salimans, “Milking cowmask for semi-supervised image classification,” *CoRR*, vol. abs/2003.12022, 2020. [Online]. Available: https://arxiv.org/abs/2003.12022

[146] Z. Zhou *et al.*, “Unet++: Redesigning skip connections to exploit multiscale features in image segmentation,” *IEEE Trans. Medical Imaging*, vol. 39, no. 6, pp. 1856–1867, 2020. [Online]. Available: https://doi.org/10.1109/TMI.2019.2959609

[147] A. Paszke *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach *et al.*, Eds., 2019, pp. 8024–8035.

[148] S. van der Walt *et al.*, “scikit-image: Image processing in python,” *CoRR*, vol. abs/1407.6245, 2014. [Online]. Available: http://arxiv.org/abs/1407.6245

[149] I. Loshchilov and F. Hutter, “SGDR: stochastic gradient descent with restarts,” *CoRR*, vol. abs/1608.03983, 2016. [Online]. Available: http://arxiv.org/abs/1608.03983

[150] P. Naylor *et al.*, “Segmentation of nuclei in histopathology images by deep regression of the distance map,” *IEEE Transactions on Medical Imaging*, vol. 38, no. 2, pp. 448–459, 2019.

[151] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1409.1556

[152] C. Szegedy *et al.*, “Rethinking the inception architecture for computer vision,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016.* IEEE Computer Society, 2016, pp. 2818–2826.

[153] G. Huang *et al.*, “Densely connected convolutional networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017.* IEEE Computer Society, 2017, pp. 2261–2269.

[154] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019.

[155] S.-A. Rebuffi, H. Bilen, and A. Vedaldi, "Learning multiple visual domains with residual adapters," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, p. 506–516.

[156] S. Rebuffi, H. Bilen, and A. Vedaldi, "Efficient parametrization of multi-domain deep neural networks," in *2018 IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8119–8127.

[157] G. Danuser, "Computer vision in cell biology," *Cell*, vol. 147, pp. 973–8, 11 2011.

[158] C. Szegedy *et al.*, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[159] E. Meijering, "A bird's-eye view of deep learning in bioimage analysis," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 2312 – 2325, 2020.

[160] M. Gadermayr *et al.*, "Domain adaptive classification for compensating variability in histopathological whole slide images," in *Image Analysis and Recognition - 13th International Conference*, vol. 9730, 2016, pp. 616–622.

[161] ——, "Which way round? A study on the performance of stain-translation for segmenting arbitrarily dyed histological images," in *Medical Image Computing and Computer Assisted Intervention*, vol. 11071, 2018, pp. 165–173.

[162] M. Lafarge *et al.*, "Learning domain-invariant representations of histological images," *Frontiers in Medicine*, vol. 6, 07 2019.

[163] A. S. Sebag *et al.*, "Multi-domain adversarial learning," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. [Online]. Available: https://openreview.net/forum?id=Sklv5iRqYX

[164] N. A. Koohbanani *et al.*, "Self-path: Self-supervision for classification of pathology images with limited annotations," *IEEE Transactions on Medical Imaging*, pp. 1–1, 2021.

[165] J. Ren *et al.*, "Adversarial domain adaptation for classification of prostate histopathology whole-slide images," in *Medical Image Computing and Computer Assisted Intervention*, vol. 11071, 2018, pp. 201–209.

[166] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations*, Y. Bengio and Y. LeCun, Eds., 2015.

[167] Y. Guo *et al.*, "Depthwise convolution is all you need for learning multiple visual domains," in *The Thirty-Third AAAI Conference on Artificial Intelligence*, 2019, pp. 8368–8375.

[168] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1800–1807.

[169] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning*, J. Fürnkranz and T. Joachims, Eds., 2010, pp. 807–814.

[170] L. Shamir *et al.*, "IICBU 2008: a proposed benchmark suite for biological image analysis," *Medical Biol. Eng. Comput.*, vol. 46, no. 9, pp. 943–947, 2008.

[171] J. Jantzen *et al.*, "Pap-smear benchmark data for pattern classification," *Nature Inspired Smart Information Systems (NiSIS)*, 01 2005.

[172] K. He *et al.*, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *2015 IEEE International Conference on Computer Vision*, 2015, pp. 1026–1034.

[173] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of adam and beyond," *CoRR*, vol. abs/1904.09237, 2019. [Online]. Available: http://arxiv.org/abs/1904.09237

[174] C. Szegedy *et al.*, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[175] Yu-Shi Lin *et al.*, "Feature space transformation for semi-supervised learning for protein subcellular localization in fluorescence microscopy images," in *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2009, pp. 414–417.

[176] L. Nanni, S. Ghidoni, and S. Brahnam, "Ensemble of convolutional neural networks for bioimage classification," *Applied Computing and Informatics*, vol. ahead-of-print, 06 2018.

[177] D. Liu *et al.*, "Medical image classification using spatial adjacent histogram based on adaptive local binary patterns," *Computers in biology and medicine*, vol. 72, pp. 185–200, 2016.

[178] D. Lin *et al.*, “Biomedical image classification based on a cascade of an svm with a reject option and subspace analysis,” *Computers in biology and medicine*, vol. 96, pp. 128–140, 2018.

[179] V. Uhlmann, S. Singh, and A. E. Carpenter, “CP-CHARM: segmentation-free image classification made accessible,” *BMC Bioinform.*, vol. 17, p. 51, 2016.

[180] L. Shamir *et al.*, “Wndchrm - an open source utility for biological image analysis,” *Source Code Biol. Medicine*, vol. 3, 2008.

[181] L. Nanni *et al.*, “Bioimage classification with handcrafted and learned features,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 3, pp. 874–885, 2019.

[182] X. Zhang and S. Zhao, “Fluorescence microscopy image classification of 2d hela cells based on the capsnet neural network,” *Medical Biol. Eng. Comput.*, vol. 57, no. 6, pp. 1187–1198, 2019. [Online]. Available: https://doi.org/10.1007/s11517-018-01946-z

[183] M. Boland et al., “A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of hela cells.” *Bioinformatics*, Dec 2001.

[184] M. Ranzato *et al.*, “Automatic recognition of biological particles in microscopic images,” *Pattern Recognition Letters*, vol. 28, no. 1, pp. 31 – 39, 2007.

[185] M. Boland et al., “Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images.” *Cytometry*, Nov 1998.

[186] N. Hamilton et al., “Fast automated cell phenotype image classification,” *BMC Bioinformatics*, vol. 8, no. 1, p. 110, Mar 2007.

[187] D. Lin *et al.*, “Llc encoded bow features and softmax regression for microscopic image classification,” in *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2017, pp. 1–4.

[188] N. e. a. Orlov, “Wnd-charm: Multi-purpose image classification using compound image transforms.” *Pattern Recognition Letters*, vol. 29, no. 11, pp. 1684–1693, 2008.

[189] T. Parnamaa, “Accurate classification of protein subcellular localization from high-throughput microscopy images using deep learning,” *G3: Genes, Genomes, Genetics*, vol. 7, no. 5, pp. 1385–1392, 2017.

[190] X. Zhang and S.-G. Zhao, "Fluorescence microscopy image classification of 2d hela cells based on the capsnet neural network," *Medical & Biological Engineering & Computing*, Jan 2019.

[191] A. G. Howard *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *CoRR*, 2017. [Online]. Available: http://arxiv.org/abs/1704.04861

[192] and J. Vieira, "2-d wavelet transforms in the form of matrices and application in compressed sensing," in *2010 8th World Congress on Intelligent Control and Automation*, July 2010, pp. 35–39.

[193] P. Thul *et al.*, "A subcellular map of the human proteome," *Science*, vol. 356, 05 2017.

[194] D. Lin *et al.*, "An SVM based scoring evaluation system for fluorescence microscopic image classification," in *2015 IEEE International Conference on Digital Signal Processing, DSP 2015, Singapore, July 21-24, 2015*. IEEE, 2015, pp. 543–547. [Online]. Available: https://doi.org/10.1109/ICDSP.2015.7251932

[195] L. Shamir *et al.*, "IICBU 2008: a proposed benchmark suite for biological image analysis," *Medical Biol. Eng. Comput.*, vol. 46, no. 9, pp. 943–947, 2008. [Online]. Available: https://doi.org/10.1007/s11517-008-0380-5

[196] J. Xie, R. B. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, ser. JMLR Workshop and Conference Proceedings, M. Balcan and K. Q. Weinberger, Eds., vol. 48. JMLR.org, 2016, pp. 478–487. [Online]. Available: http://proceedings.mlr.press/v48/xieb16.html

[197] J. Macqueen, "Some methods for classification and analysis of multivariate observations," in *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281–297.

[198] H. W. Kuhn, "The hungarian method for the assignment problem," in *50 Years of Integer Programming 1958-2008 - From the Early Years to the State-of-the-Art*, M. Jünger *et al.*, Eds. Springer, 2010, pp. 29–47. [Online]. Available: https://doi.org/10.1007/978-3-540-68279-0_2

[199] M. Majurski *et al.*, "Cell image segmentation using generative adversarial networks, transfer learning, and augmentations," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June*

16-20, 2019. Computer Vision Foundation / IEEE, 2019, pp. 1114–1122. [Online]. Available: http://openaccess.thecvf.com/content_CVPRW_2019/html/CVMI/Majurski_Cell_Image_Segmentation_Using_Generative_Adversarial_Networks_Transfer_Learning_and_CVPRW_2019_paper.html

[200] S. Zhao, J. Song, and S. Ermon, "Infovae: Balancing learning and inference in variational autoencoders," in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019.* AAAI Press, 2019, pp. 5885–5892. [Online]. Available: https://doi.org/10.1609/aaai.v33i01.33015885

[201] T. Veniat, L. Denoyer, and M. Ranzato, "Efficient continual learning with modular networks and task-driven priors," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021.* OpenReview.net, 2021. [Online]. Available: https://openreview.net/forum?id=EKV158tSfwv

[202] D. A. Forsyth and J. Ponce, *Computer Vision - A Modern Approach, Second Edition*, 2012.

[203] X. Yang *et al.*, "Deep spectral clustering using dual autoencoder network," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2019, pp. 4066–4075.

[204] J. Macqueen, "Some methods for classification and analysis of multivariate observations," in *In 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281–297.

[205] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions On Pattern Analysis And Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[206] D. Cai and X. Chen, "Large scale spectral clustering via landmark-based sparse representation," *IEEE Transactions on Cybernetics*, vol. 45, no. 8, pp. 1669–1680, Aug 2015.

[207] J. Xie, R. B. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proceedings of the 33nd International Conference on Machine Learning, ICML*, vol. 48, 2016, pp. 478–487.

[208] N. Dilokthanakul *et al.*, “Deep unsupervised clustering with gaussian mixture variational autoencoders,” *CoRR*, vol. abs/1611.02648, 2016. [Online]. Available: http://arxiv.org/abs/1611.02648

[209] K. G. Dizaji *et al.*, “Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization,” in *IEEE International Conference on Computer Vision, ICCV*, 2017, pp. 5747–5756.

[210] Z. Jiang *et al.*, “Variational deep embedding: An unsupervised and generative approach to clustering,” in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI*, 2017, pp. 1965–1972.

[211] S. Zhao, J. Song, and S. Ermon, “Infovae: Balancing learning and inference in variational autoencoders,” in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI,*, 2019, pp. 5885–5892.

[212] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2014. [Online]. Available: http://arxiv.org/abs/1312.6114

[213] X. Guo *et al.*, “Improved deep embedded clustering with local structure preservation,” in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI*, 2017, pp. 1753–1759.

[214] ——, “Deep clustering with convolutional autoencoders,” in *Neural Information Processing - 24th International Conference, ICONIP*, vol. 10635, 2017, pp. 373–382.

[215] J. Yang, D. Parikh, and D. Batra, “Joint unsupervised learning of deep representations and image clusters,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2016, pp. 5147–5156.

[216] I. J. Goodfellow *et al.*, “Generative adversarial nets,” in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems*, 2014, pp. 2672–2680.

[217] S. Mukherjee *et al.*, “Clustergan: Latent space clustering in generative adversarial networks,” in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI*, 2019, pp. 4610–4617.

[218] A. Gretton *et al.*, “A kernel method for the two-sample-problem,” in *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems,*, B. Schölkopf, J. C. Platt, and T. Hofmann, Eds., 2006, pp. 513–520.

[219] L. van der Maaten and G. Hinton, “Visualizing data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[220] Y. Le Cun *et al.*, “Handwritten zip code recognition with multilayer networks,” in *[1990] Proceedings. 10th International Conference on Pattern Recognition*, vol. ii, 1990, pp. 35–40 vol.2.

[221] H. Xiao, K. Rasul, and R. Vollgraf, “Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms,” *CoRR*, vol. abs/1708.07747, 2017.

[222] A. Krizhevsky and G. Hinton, “Learning multiple layers of features from tiny images,” *Computer Science Department, University of Toronto, Tech. Rep*, vol. 1, 01 2009.

[223] H. W. Kuhn, “The hungarian method for the assignment problem,” in *50 Years of Integer Programming 1958-2008 - From the Early Years to the State-of-the-Art.* Springer, 2010, pp. 29–47.

[224] M. Caron *et al.*, “Deep clustering for unsupervised learning of visual features,” in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV*, ser. Lecture Notes in Computer Science, V. Ferrari *et al.*, Eds., vol. 11218. Springer, 2018, pp. 139–156.

[225] X. Chen *et al.*, “Infogan: Interpretable representation learning by information maximizing generative adversarial nets,” in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems*, 2016, pp. 2172–2180.

APPENDIX A ARTICLE 5: DEEP IMAGE CLUSTERING USING SELF-LEARNING OPTIMIZATION IN A VARIATIONAL AUTO-ENCODER

Duc Hoa Tran[1], Michel Meunier[2], Farida Cheriet[1]

[1]Department of Computer and Software Engineering, Polytechnique Montréal, Canada
[2]Department of Engineering Physics, Polytechnique Montréal, Canada

Presentation

This work was published in the conference proceedings of "The third International Workshop on Deep Learning for Pattern Recognition" on February 23, 2021 [128].

Abstract

Deep image clustering approaches typically use autoencoder architectures to learn compressed latent representations suitable for clustering tasks. However, they do not effectively regulate the latent space during training, leading to low performance and diminished applicability to different datasets. In this paper, we propose a deep clustering model combining maximum mean discrepancy (MMD) regularization and self-learning clustering optimization to mitigate this problem. Specifically, we first train the network to improve its image reconstruction ability by minimizing both reconstruction loss and MMD divergence from a target distribution. Then, the model gradually learns from its own high-confidence predictions to further optimize the latent distribution. We validate the network's performance on different benchmark image sets using standard clustering metrics, without changing network configuration or adjusting hyper-parameters between datasets. The proposed model provides top clustering performance across datasets while being more robust than state-of-the-art methods.

Introduction

Clustering is an unsupervised learning approach that is essential for image analysis tasks, especially image categorization and segmentation [202], [203]. Conventional methods like K-means [204] and spectral clustering [205], [206] have been applied to a wide range of applications. However, as they measure the similarity distance between points in shallow, high-dimensional feature spaces of raw pixels or gradient-based histograms, their clustering ability is largely limited to simple image datasets [207].

For their part, deep learning-based clustering approaches use deep neural networks to represent data as lower-dimension, hierarchical features such that conventional clustering techniques can be applied effectively [208]. For example, an encoder-decoder architecture can be used to produce a rich latent encoding of the input image with dramatically reduced dimensionality. Thus, a data grouping applying on the latent codes is more feasible than performing on the input images and can avoid the curse of dimensionality problem to which clustering algorithms are prone.

Typically, there are two major challenges when applying existing deep clustering algorithms to different image datasets: (1) maintaining high algorithm performance requires reconfiguring the network architecture or adjusting a large number of training hyper-parameters [209]; (2) complicated algorithms such as spectral clustering can achieve very high performance but dramatically increase model complexity and memory usage due to the need for extensive computations, e.g. computing the full graph Laplacian matrix [207]. Therefore, our aim is on designing an algorithm that has a limited number of hyper-parameters and computational complexity but still reaches top-level performance on several different image sets without the need for reconfiguring the network architecture.

In this paper, we present a deep clustering model based on a generative Variational Autoencoder (VAE) architecture, or MMD-VAE based **D**eep **E**mbedded **C**lustering, denoted as MMV-DEC, and devise training strategies to improve its clustering performance. Unlike conventional clustering algorithms or dimension reduction techniques, which use linear transformation, our method can perform complex non-linear transformations using a deep convolutional neural network (CNN). Our work differs from previous related works, especially another VAE-based approach in [210], in terms of architecture, optimization and performance. Firstly, instead of linear layers, we use convolutional layers to improve the feature extraction capability on image data. Secondly, inspired by a recent published work [211], we use the MMD divergence optimization approach instead of maximizing the Evidence Lower Bound Objective (ELBO) based on Kullback-Leibler (KL) divergence [212]. This helps to avoid the problem of vanishing mutual information between the input image and the embedded latent code. Finally, we integrate a self-learning optimization technique to improve the clustering quality. We demonstrate a significant improvement in performance and generalizability compared with state-of-the-art models by experiments on four different benchmark datasets.

Related Works

Deep learning-based clustering has been widely studied in recent years. An early work presented in [207] proposed to use a fully connected stacked autoencoder architecture, with a two-phase training strategy. In the first phase, the network learns a feature transformation via an image reconstruction task. Then in the second phase, a clustering objective is defined based on the network's own predictions to further optimize its parameters and the cluster centroids. Improvements to this architecture have been proposed, with the reconstruction task is maintained during the second phase to preserve the structure of the data generating distribution [213] and convolutional layers are used instead of fully connected layers [214]. A

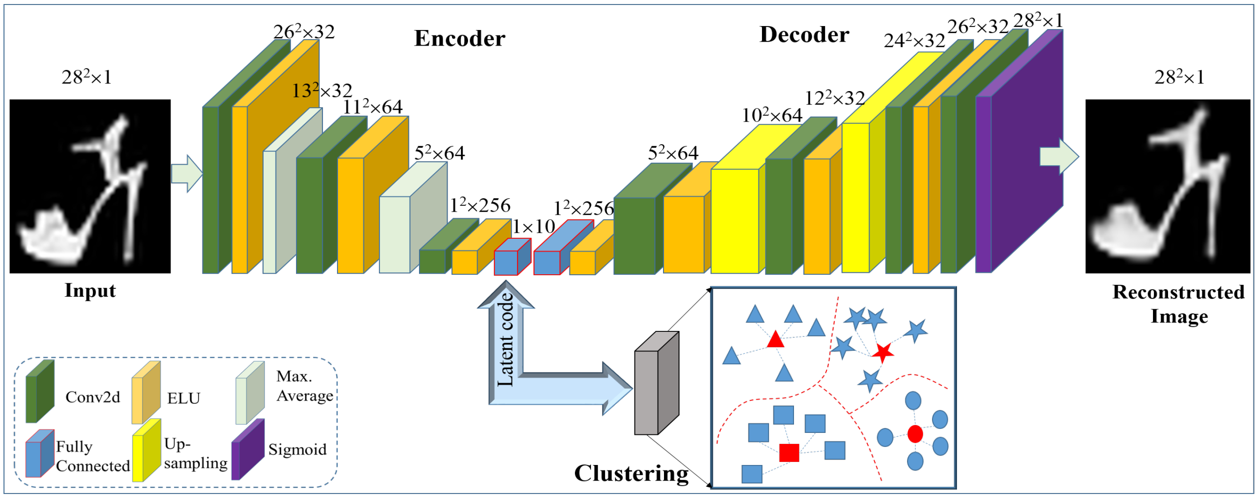


Figure A.1 Overview of our proposed model, which consists of encoder-decoder pathways and a clustering layer stacked on top of the latent layer. The dimensions shown above the layers are those of the feature maps generated at each layer.

more effective approach was introduced in [209], based on jointly and simultaneously optimizing a clustering objective and the autoencoder parameters, without layer-wise pretraining of the autoencoder layers. To achieve noise-invariant predictions, the clustering loss objective is applied on the latent code of a denoising convolutional auto-encoder, whereas the reconstruction loss function is calculated between all the decoder layers and the clean encoder layers. Recently, the authors of [203] proposed to combine spectral clustering with a dual autoencoder, with one encoder pathway for clean input and the other for its noise-contaminated version. To learn more discriminative information from the inputs, they maximize mutual information calculated with a negative image sample randomly selected from the noisy batch. After training this autoencoder for initial latent representation, the latent representations are embedded into the eigenspace of their associated graph Laplacian matrix where clustering is performed. As an alternative training strategy, Joint Unsupervised Learning (JULE) [215] combines the feature representation of a CNN and agglomerative clustering in a recurrent manner. The algorithm starts with an initial over-clustering and alternates between two training steps: merging clusters based on the current network representation and updating network parameters using the current clustering result.

Variational Autoencoders (VAE) [212] and Generative Adversarial Networks (GAN) [216] also have an encoder-decoder network structure and are popular choices for modeling the process of data generation and synthesizing new images. Although VAEs and GANs yield better reconstruction performance in various applications, there is a limited number of published deep clustering algorithms based on VAE or GAN architectures. A clustering approach based on a VAE was introduced in [210]. To be more pertinent for clustering tasks, it models the data generation process by a Mixture of Gaussian prior instead of the original Gaussian prior. The model is optimized using the conventional method of maximizing the ELBO of the data log-likelihood as well as the re-parameterization trick. Another recent work exploits a generative model that performs latent space clustering in a GAN [217]. As the cluster structure is not held in the GAN latent space, they propose to use the mixture sampling of discrete and continuous latent variables and a set of optimization algorithms specialized for the discrete-continuous mixture.

Proposed Approach

Our deep clustering model, called MMV-DEC, is composed of two main parts: a VAE network based on MMD regularization, which we name MVAE and a clustering layer which is stacked on top of the latent layer of the MVAE to enable enhanced clustering (EC) optimization. The overall architecture is shown in Fig. A.1.

MVAE

The proposed variational autoencoder network consists of two major components, an encoder and a decoder, each comprising a set of convolutional layers. The encoder extracts the input image features to produce the latent code, which has much lower dimension than the input image. This latent code is fed into the decoder layers to reconstruct the original image. The network functions as a generative model, with encoder g_θ and decoder f_ϕ being functions of the network parameter sets θ and ϕ:

$$X \xrightarrow{g_\theta} Z \xrightarrow{f_\phi} \hat{X}$$

We design the encoder network using three convolutional layers, with 32 kernels of size 3×3, 64 kernels of size 3×3 and 256 kernels of size 5×5, used in each layer, respectively. Each of these convolutional layers is followed by an Exponential Linear Unit (ELU) activation function. In addition, average pooling with kernel size 2×2 and stride of 2 pixels is applied on each of the activated feature maps, except for the third feature map. The pooling operation functions as a down-sampling layer to reduce the spatial dimension. After the third feature map, a fully connected layer is used to generate the latent vector of size 1×10 corresponding to the original grey-level image of size 28×28.

Conversely, the decoder component is designed to gradually increase the feature map dimensions, starting from the latent vector and ending with the output reconstructed image. The decoder pathway starts with one fully connected layer which is followed by four convolutional layers integrated with ELU activation functions. Each of the four convolutional layers uses a set of 64 kernels of size 5 × 5, 32 kernels of size 3 × 3, 16 kernels of size 3 × 3 and 1 kernel of size 3 × 3, respectively. In addition, appropriate padding is used in combination with bilinear up-sampling layers after each activated feature map such that the reconstructed images have the same size as the original images. The last layer uses a conventional sigmoid function to normalize the output values in the range of $[0, 1]$ to produce the output image.

Typically, the cost function to train a VAE includes two terms: the reconstruction loss and the regularization loss. The main difference between a standard VAE and our MVAE network lies in the formulation of the regularization loss. Whereas conventional VAE optimization is based on minimizing the Kullback-Leibler divergence between the generated latent distribution and a prior distribution, we apply Maximum Mean Discrepancy (MMD) divergence [211] to optimize jointly with the reconstruction loss.

During training, each unlabeled image is fed into the network and the reconstruction loss, known as the binary cross entropy (BCE) function, is calculated. Supposing that in each training iteration, a batch of b images is processed, then the BCE loss between input images x_i of m pixels and their reconstructed counterparts $\hat{x}_i$ is measured element-wise by:

$$BCE = -\frac{1}{b \times m} \sum_{i=1}^{b} \sum_{j=1}^{m} [x_{ij} \ln \hat{x}_{ij} + (1 - x_{ij})(1 - \ln(1 - \hat{x}_{ij}))] \tag{A.1}$$

Simultaneously, the MMD divergence between the distribution of generated latent variables $q(z)$ and a target distribution $p(z')$ is computed by using the kernel embedding formula [211], [218]:

$$MMD = E_{p(z'_i),p(z'_j)}[k(z'_i, z'_j)] + E_{q(z_i),q(z_j)}[k(z_i, z_j)] - 2E_{p(z'_i),q(z_j)}[k(z'_i, z'_j)] \tag{A.2}$$

where

$$k(z_i, z_j) = e^{-\frac{||z_i - z_j||}{2\sigma^2}}$$

is a kernel to measure the similarity between two samples z_i, z_j in terms of Euclidean distance. Here, $q(z)$ is the distribution of latent variables generated by the encoder, while $p(z')$ is the prior distribution that we would like $q(z)$ to match. Intuitively, the MMD loss measures the difference between the average similarity of samples within each distribution and the average similarity of mixed samples from both distributions. When MMD reaches 0, the two distributions are matched. Finally, the model is trained to optimize the aggregate cost function, which is equal to the sum of the reconstruction and MMD losses:

$$L_1 = BCE + MMD \tag{A.3}$$

Enhanced clustering optimization (EC)

To further optimize the latent space for clustering purposes, we borrow the unsupervised self-learning optimization technique proposed in [207]. To achieve this, we connect an additional clustering layer to the latent variable layer, leaving the rest of the MVAE network intact. It is a fully connected layer that uses a t-distribution kernel [219] to measure the similarity between the latent code and the centroid of a target cluster. Specifically, the distance between a sample z_i and the centroid of a given cluster μ_j is calculated as:

$$q_{ij} = \frac{(1 + ||z_i - \mu_j||^2)^{-1}}{\sum_{j'}(1 + ||z_i - \mu_{j'}||^2)^{-1}} \tag{A.4}$$

This equation calculates the probability of a data point belonging to a cluster represented by its mean μ_j and is translated into the assignment of a class label to the input image. Note that standard K-means clustering is applied in the latent space to determine the initial clusters. Then, during the optimization process, the cluster centroids are updated as learnable parameters.

The self-learning optimization involves defining a target distribution p_{ij} and minimizing the Kullback-Leibler (KL) divergence between p_{ij} and the embedding distribution calculated in equation A.4. We use a simple and effective empirical target distribution [207], defined as:

$$p_{ij} = \frac{q_{ij}^2 / \sum_i q_{ij}}{\sum_{j'}(q_{ij'}^2 / \sum_i q_{ij'})} \tag{A.5}$$

The KL divergence used to evaluate the matching between the target distribution and the clustering assignment of latent variables is computed by:

$$D_{KL}(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{A.6}$$

This KL divergence is used as a regularization term and serves as a guidance criterion for refining the clusters. As p_{ij} is also a function of q_{ij}, optimizing this divergence is considered a self-training process. However, instead of using this single KL divergence as the cost function for optimization, we combine it with a reconstruction loss as suggested in [213] to preserve the local structure of the data distribution and avoid overfitting or getting stuck in local minima [209] during network optimization. Supposing that in each training iteration, a batch of b images is processed, then the element-wise mean squared error (MSE) between b

input images x_i and their reconstructed images $\hat{x}_i$ is measured by:

$$MSE = \frac{1}{b \times m} \sum_{i=1}^{b} ||x_i - \hat{x}_i||_2^2 \tag{A.7}$$

So the final cost function for enhanced clustering optimization is formulated as the weighted sum of the KL divergence and reconstruction loss:

$$L_2 = MSE + \beta D_{KL} \tag{A.8}$$

where β is a hyper-parameter to adjust the weight of KL divergence and MSE is the reconstruction loss.

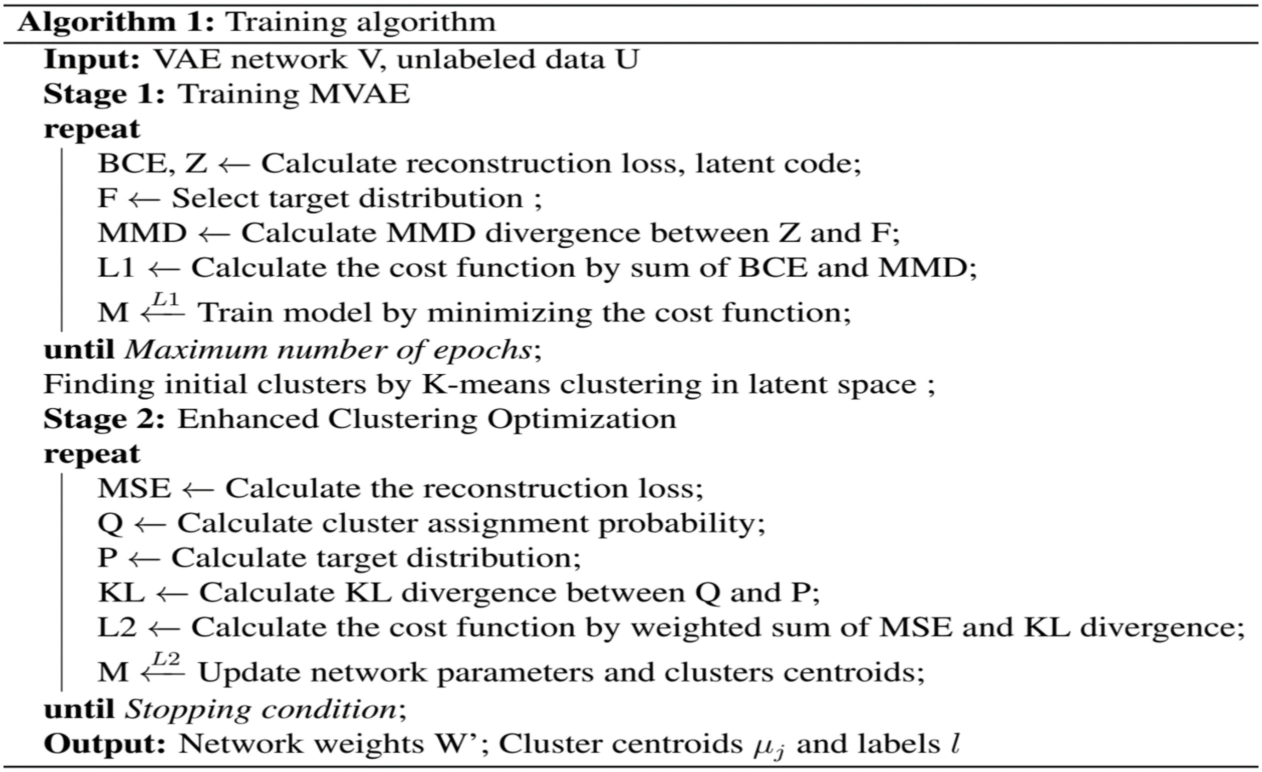
Training

We use the training algorithm presented in Fig. A.2 for all datasets, with the same network configuration and a fixed set of hyper-parameters. It is divided into two major stages: (1) training the MVAE network as a generative model and (2) training the network for enhanced clustering optimization. In the first stage, the training process uses standard backpropagation to update parameters, together with ADAM optimization at a fixed learning rate of 0.001. We select the Gaussian distribution $p(z') \sim \mathcal{N}(0, 0.5)$ as the target distribution for embedding variables. The maximum number of training epochs is set to 200 in order for the reconstructed images to achieve relatively good quality for all the tested datasets. Then, we use simple K-means clustering on the latent variables to find the initial cluster centroids, which are required for equation A.4 in the first iteration of the clustering optimization stage. In the second stage, we again use the ADAM optimizer with a fixed learning rate of 0.001 to update the network parameters, including the cluster centroids, after every training iteration. Similar to [213], we set $\beta = 0.1$ to balance the contributions of the loss terms in the L_2 cost function. The stopping condition is triggered when the difference in cluster assignments compared with the previous iteration is below a small threshold.

Experiments and discussion

Datasets and evaluation metrics

To compare our performance results with recently published methods, we evaluated our proposed model on 4 reference image sets, namely MNIST [5], USPS [220], Fashion-MNIST [221], and Cifar-10 [222] as summarized in Table A.1. For the Cifar-10 dataset which consists of color images, we adjust the dimension of the input tensor accordingly. Two standard unsupervised evaluation metrics for clustering performance were used: clustering Accuracy (ACC) and Normalized Mutual Information (NMI) [203], [217]. The classification label for each input image was produced by applying the well known Hungarian algorithm [223] which



```
Algorithm 1: Training algorithm
  Input: VAE network V, unlabeled data U
  Stage 1: Training MVAE
  repeat
      BCE, Z ← Calculate reconstruction loss, latent code;
      F ← Select target distribution ;
      MMD ← Calculate MMD divergence between Z and F;
      L1 ← Calculate the cost function by sum of BCE and MMD;
      M $\xleftarrow{L1}$ Train model by minimizing the cost function;
  until Maximum number of epochs;
  Finding initial clusters by K-means clustering in latent space ;
  Stage 2: Enhanced Clustering Optimization
  repeat
      MSE ← Calculate the reconstruction loss;
      Q ← Calculate cluster assignment probability;
      P ← Calculate target distribution;
      KL ← Calculate KL divergence between Q and P;
      L2 ← Calculate the cost function by weighted sum of MSE and KL divergence;
      M $\xleftarrow{L2}$ Update network parameters and clusters centroids;
  until Stopping condition;
  Output: Network weights W'; Cluster centroids $\mu_j$ and labels $l$
```

Figure A.2 Training algorithm for MMV-DEC.

maps the predicted clusters assignments to the groundtruth labels. To reduce measurement uncertainty, the performance measurements were averaged from 10 random trials.

Image reconstruction

The MVAE embeds the input image into a low-dimensional latent code by the encoder layers and then reconstructs the original image by the decoder layers. By minimizing the integrated reconstruction loss and MMD regularization term, the network is supposed to learn the underlying data representation more effectively than a conventional autoencoder or VAE and

Table A.1 Image datasets for clustering evaluation

Dataset	# Images	# Classes	Dimension
MNIST	70,000	10	$28 \times 28 \times 1$
USPS	9,298	10	$16 \times 16 \times 1$
Fashion-MNIST	70,000	10	$28 \times 28 \times 1$
Cifar-10	60,000	10	$32 \times 32 \times 3$

thus produce better reconstruction quality. Examples of original and reconstructed images obtained by our model are illustrated in Fig. A.3.

Even though the latent representation is significantly compressed compared with input dimensions, in the three datasets MNIST, USPS and Fashion-MNIST, the reconstructed images produced by the decoder (in the even rows) are visually very close to the original ones (in the odd rows), with some minor blurring. Therefore, the useful information to discriminate differing patterns is maintained in the latent code, which is a desired condition for clustering to be directly performed on it. On the other hand, the reconstructed images for Cifar-10 dataset are very blurry which implies the necessity of the preprocessing step and scaling up of the deep CNN architecture. For example, the authors of [224] used Sobel transformation to convert the color images into grayscale images and employed AlexNet or VGG networks for their feature extraction.

Analysis of training strategies

We validated the effectiveness of applying the MMD regularization and enhanced clustering (EC) optimization strategies by comparing five variants of our model: (1) Convolutional autoencoder using our encoder-decoder architecture but trained with only the reconstruction loss (ConvAE); (2) Convolutional autoencoder using enhanced clustering optimization (ConvAE+EC); (3) Convolutional autoencoder trained with both reconstruction and conventional KL divergence losses (ConvAE+KL div.); (4) Convolutional autoencoder trained with both reconstruction and MMD losses (ConvAE+MMD), which is the MVAE network; (5) MVAE network using enhanced clustering optimization (ConvAE +MMD+EC), which is our proposed MMV-DEC model.

As shown in Table A.2, each training strategy of MMD loss and EC improves the clustering ACC and NMI results effectively on all four benchmarks, especially on the MNIST and USPS datasets. The application of both techniques consistently produces the highest performance, thus demonstrating the appropriateness of combining them in training. Although the improvement is not significant in the case of the Fashion-MNIST dataset and Cifar-10, the relatively high performance of the basic configuration (ConvAE) implies that the network architecture is well designed, laying the foundation for other optimization strategies. As can also be seen in the table, the application of conventional KL divergence loss during the network training is inferior to MMD loss optimization and even deteriorate the clustering performance of the autoencoder network in case of the Fashion-MNIST dataset.

Fig. A.4 provides a visual comparison of the latent space for the USPS dataset by applying the t-SNE visualization method [219] on the embedded code space Z. This visualization reveals that training with the combined reconstruction and MMD losses (MVAE network)

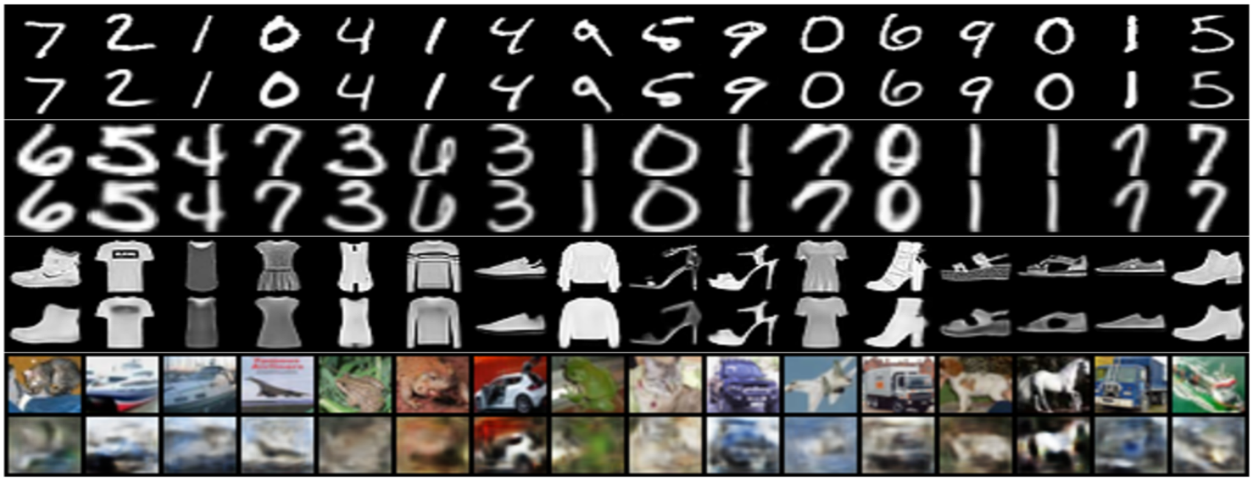


Figure A.3 Original and reconstructed images from the three different datasets: MNIST (rows 1 & 2), USPS (rows 3 & 4), Fashion-MNIST (rows 5 & 6) and Cifar-10 (rows 7 & 8).

produces more compact clusters that are easier to discriminate, compared with using only the reconstruction loss (ConvAE network).

Fig. A.5 displays the improvement in clustering accuracy and NMI during Stage 2 of training (EC optimization) of our network for one trial example. We can see here that the clustering optimization technique is generally very fast and effective. In addition, the output performance of the Stage 1-trained MVAE plays a vital role in laying the basis for further clustering optimization. Indeed, the clustering optimization increases performance significantly on the MNIST (ACC: $\sim 7\%$, NMI: $\sim 12\%$) and USPS (ACC: $\sim 7\%$, NMI: $\sim 11\%$) datasets; in these cases, the Stage 1-trained MVAE provides high clustering capability to initiate Stage 2. For the Fashion-MNIST and Cifar-10 datasets, however, the improvement is more limited, as it is impacted by the quality of the previously trained MVAE.

Comparison with state of the art methods

We compared our proposed approach with both conventional clustering baselines and state-of-the-art deep clustering algorithms. Peformance results for these other methods were reported either in their original papers or compiled in recently published papers [203], [217]. For the Cifar-10 dataset, we obtain the results by running the released codes of corresponding works and for those results that are not practical to obtain, we indicate by dash marks ($-$). As can be seen in Table A.3, our proposed approach outperforms the conventional methods based on K-means or Spectral clustering (SC-LS) on all four datasets by a large margin. Furthermore, our model yields better performances than several other deep embedded clustering methods, including DEC, IDEC, DCEC, as well as a generative model based on the variational au-

Table A.2 Analysis of different training strategies

Strategies	**MNIST**		**USPS**		**Fashion**		**Cifar-10**	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
ConvAE	86.6	77.7	72.3	70.9	61.5	64.0	21.9	8.9
ConvAE+EC	95.2	91.2	78.7	82.7	62.1	65.8	23.3.0	9.5
ConvAE+KL div.	86.9	79.7	78.2	72.5	50.4	48.2	22.8	9.1
ConvAE+MMD	90.0	81.0	85.9	78.6	61.7	64.9	23.0	9.8
Our MMV-DEC	**96.8**	**93.3**	**96.4**	**91.2**	**62.9**	**66.2**	**24.1**	**10.4**

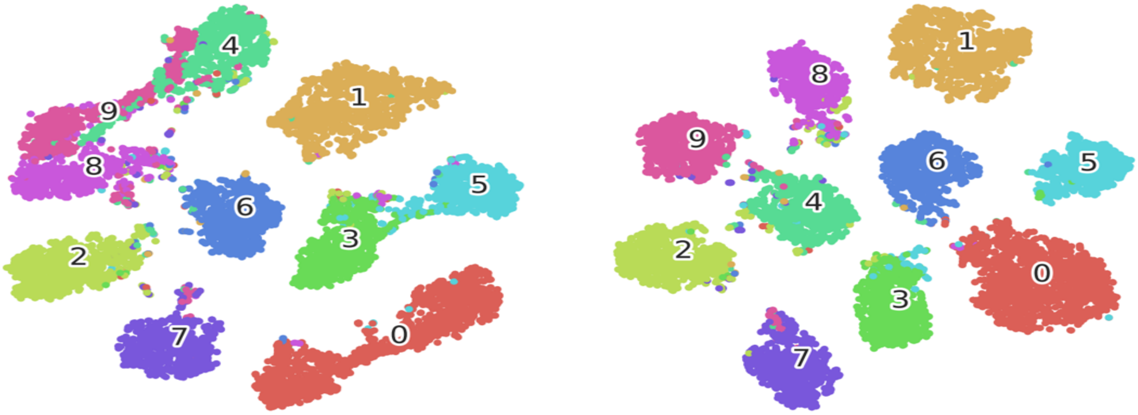


Figure A.4 Comparison of latent representations of USPS dataset produced by our model without (left) and with (right) MMD regularization. Colors represent true labels of the samples, and class numbers are positioned at cluster centroids.

toencoder (VaDE). Compared with the methods achieving the highest performances in the literature, namely JULE, DEPICT, Dual-AE and the GAN models (ClusterGAN, InfoGAN), our proposed MMV-DEC achieves higher overall clustering performance. In particular, it outperforms other methods on Cifar-10 dataset following both ACC and NMI criteria. It also achieves the highest accuracy (96.4%) on the USPS and the highest NMI (66.2%) on the Fashion-MNIST dataset. In general, our MMV-DEC secures at least 2nd best performance according to the two metrics on all benchmarks.

More importantly, the experiment results demonstrate the better generalization ability across datasets of our proposed framework. Note that none of the previous methods gain a top-

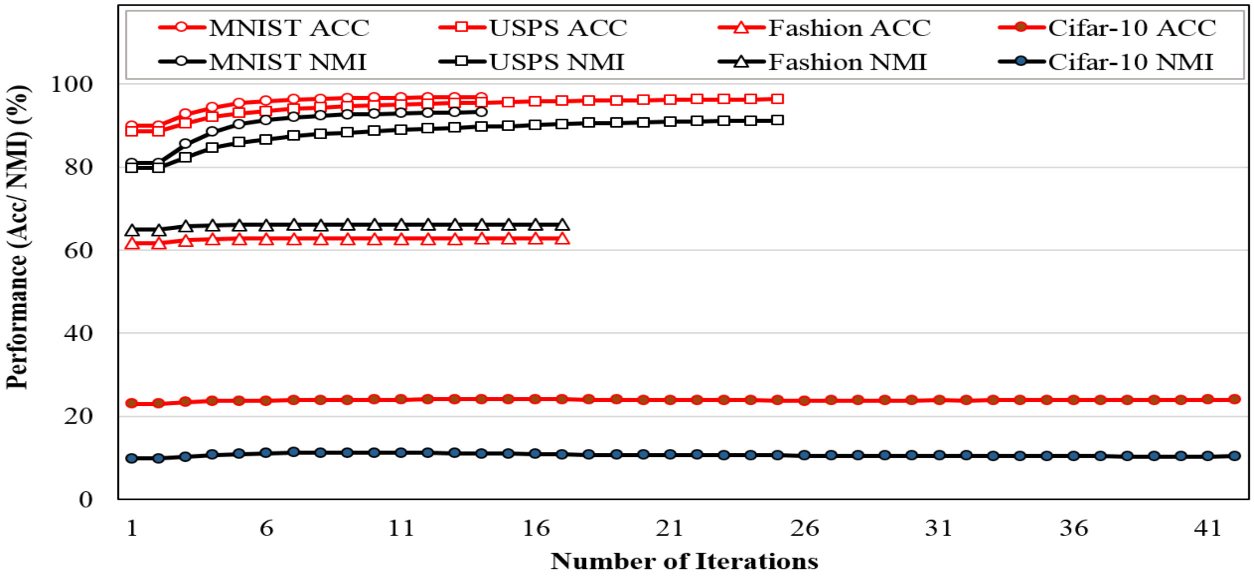


Figure A.5 ACC and NMI metrics during EC optimization (Stage 2 of training) with MVAE network.

two performance across all datasets and the latest state of the art methods could not provide consistent results. For example, although the USPS and MNIST datasets are similar, the difference in clustering accuracy is more than 10% in most of recent works [203,207,213,214] and it is even more than 35% for the similar approach based on conventional VAE network [210]. With our proposed method, this performance gap is non-remarkable and without the need of adjusting the training hyper-parameters.

Conclusion

In this paper, we present a new unsupervised deep clustering method that is based on a variational autoencoder architecture. The application of self-learning mechanism and MMD loss optimization consistently produces the highest effectiveness and generalization ability. The model also has the advantages of low computational complexity and few hyper-parameters to adjust. Experiments on four image benchmarks demonstrate that our proposed MMV-DEC model can reach state-of-the-art performance without requiring to reconfigure the network architecture or change the clustering hyper-parameters. Our further work will focus on more realistic images, where it is necessary to scale up the deep network architecture and apply preprocessing steps on the input images.

Table A.3 Comparison of different clustering algorithms on benchmark datasets based on NMI and ACC metrics. The top two performances are highlighted in each column.

Methods	**MNIST**		**USPS**		**Fashion**		**Cifar-10**	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
K-means [204]	53.2	50.0	66.8	60.1	47.4	51.2	19.8	7.6
SC-LS [206]	71.4	70.6	74.6	75.5	49.6	49.7	20.6	9.1
DEC [207]	86.3	83.4	76.2	76.7	51.8	54.6	21.6	8.4
JULE [215]	96.4	91.3	**95.0**	**91.3**	56.3	60.8	-	-
VaDE [210]	94.5	87.6	56.6	51.2	57.8	63.0	20.1	8.1
IDEC [213]	88.1	86.7	76.1	78.5	52.9	55.7	19.6	8.3
DCEC [214]	89.0	88.5	79.0	82.57	-	-	22.3	8.7
DEPICT [209]	96.5	91.7	89.9	90.6	39.2	39.2	22.8	9.6
InfoGAN [225]	89.0	86.0	-	-	61.0	59.0	-	-
ClusterGAN [217]	95.0	89.0	-	-	**63.0**	64.0	-	-
Dual-AE [203]	**97.8**	**94.1**	86.9	85.7	**66.2**	**64.5**	**23.9**	**9.8**
Our MMV-DEC	**96.8**	**93.3**	**96.4**	**91.2**	62.9	**66.2**	**24.1**	**10.4**