

Advanced Learning-based Visual Content Analysis for Computer Vision

VEENA MURTHY SRINIVASA DODBALLAPUR

SID: 470430501

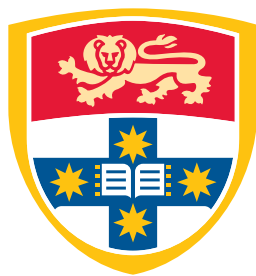
Supervisor: A/Prof. Weidong Cai

This thesis is submitted in fulfillment of
the requirements for the degree of
Doctor of Philosophy

School of Computer Science, Faculty of Engineering
The University of Sydney
Australia

This research reported in this thesis was supported by the award of a Research Training Program scholarship to the PhD Candidate.

9 April 2024



THE UNIVERSITY OF
SYDNEY

Student Plagiarism: Compliance Statement

I certify that:

I have read and understood the University of Sydney Student Plagiarism: Coursework Policy and Procedure;

I understand that failure to comply with the Student Plagiarism: Coursework Policy and Procedure can lead to the University commencing proceedings against me for potential student misconduct under Chapter 8 of the University of Sydney By-Law 1999 (as amended);

This Work is substantially my own, and to the extent that any part of this Work is not my own I have indicated that it is not my own by Acknowledging the Source of that part or those parts of the Work.

Name: Veena Murthy Srinivasa Dodballapur

Signature:

Date:

Abstract

This thesis tackles the challenges in deep neural networks by introducing advanced learning-based methods for visual content analysis tasks for computer vision. Historically, there has always been an interest to understand digital image or sequence of images to draw out useful information to aid in decision making for various real-world computer vision applications. The earlier technologies focussed on using hand-crafted features and this progressed to more robust features using deep neural networks (DNN). The DNN network has shown significant improvement in computer vision applications such as object detection and classification, counting and segmentation. The success of deep neural network models depends on good quality data encompassing different conditions such as backgrounds, occlusions, cameras, poses, morphological features and accurate annotations. However, we see that it is challenging, expensive and tedious to obtain annotated datasets which cover diverse conditions. While some applications can get-by with the coarse bounding-box annotations, pixel level annotation plays an important part in object detection accuracy, so that the noise from the background can be filtered out.

The quest for improved deep learning solutions is ongoing. Computer vision applications use techniques such as classification, object detection, segmentation and counting both widely and extensively. However, these applications have challenges such as domain variability, poor generalization, inadequate localization, and incomplete and weakly annotated datasets which need to be addressed to obtain better application accuracy. Hence, there is a need for advanced solutions which address these challenges. In the thesis, which we have divided into two parts, we tackle these challenges within both fully supervised and weakly supervised settings, contextualizing them in various applications.

In the first part of our thesis, we develop sophisticated deep learning techniques for computer vision applicable to a fully supervised context. We present methods aimed at enhancing model generalization and refining object localization. We have studied the use of these advanced deep learning methods in classification and object detection to demonstrate the effectiveness of our algorithms.

We address the challenge of poor generalization using automatic and adaptive dropout for object classification scenarios. Existing regularization methods using dropout, use hyper-parameters to specify a dropout rate. Alternatively, they rely on structural changes to the network, addition of noise, ensemble of networks or sparsity of weights. These alternatives are more expensive to implement and cannot be easily adapted in existing deep learning models. We propose two novel and improved dropout methods, which provide better generalization of deep learning models without requiring any hyper-parameters. Our methods provide an automatic and adaptive approach of determining a dropout rate using the distribution of the neuron activations, and achieves better performance across popular datasets and deep learning networks. Our method can be used as a drop-in replacement for any deep neural network with standard dropout and does not rely on structural changes to the network.

Our second method proposes a dual-stage deep learning framework based on Mask R-CNN for better localization of objects, thus improving detection accuracy. Furthermore, we extend our method to automatically generate pixel-level masks for weakly annotated datasets. We demonstrate the advantage of our method on two cell datasets with morphological variations, by improving the detection accuracy.

The fully supervised approaches provide promising results, but rely on fully annotated datasets. Training robust deep learning models require datasets that encompass all types of variations in data, including pixel-level annotations. Preparing and obtaining such datasets is expensive and impractical. Weakly supervised methods provide an advantage over fully supervised methods to address the lack of full annotation. In the second part of our thesis, we extend our fully supervised approaches to handle incomplete datasets where the major challenges are domain gap and weak annotations, and apply it to object detection, person re-identification, and cell counting.

Domain gap is a major challenge to overcome in both object detection and person identification scenarios. In the case of object detection, we build on the approach developed in the first part of our thesis to solve both domain gap and localization challenges. In this approach, we apply adversarial training strategy at image and instance level to a dual-stage domain adaptive deep neural network to overcome the domain shift caused by different staining conditions in histopathology images. Furthermore, we extend our method to predict accurate pixel-level

masks in a cross-domain setting for weakly annotated datasets. We demonstrate improved detection accuracy for datasets with small cells, achieved through better instance localization, using the masks generated in a cross-domain setting.

In yet another approach, staying on the theme of domain variance, we develop a self-supervised domain adaptive method for generating robust pseudo-labels and apply it to person re-identification. Our approach uses region-specific spatio-temporal features to generate robust pseudo-labels, which are used for self-supervised training using a dataset with no annotations. We demonstrate the robustness of our approach in handling challenges such as pose, camera and background variability between different domains, by achieving state-of-the-art results.

Finally, we build on the approaches developed in the first part of the thesis (which dealt with fully annotated datasets), to tackle weakly annotated datasets. We consider the cell counting scenario for this. Weak annotation typically occurs in counting scenarios, as there are numerous objects to count and is difficult to precisely annotate all occurrences. We study cell counting using weakly supervised learning, by automatically generating and selecting good quality exemplars using a novel cross-domain training strategy. We exploit the cell structure similarity to achieve cross-domain knowledge transfer, to generate exemplars automatically, thus eliminating the need for manual annotation by a human. To the best of our knowledge, this is the first time a method has been proposed to define exemplars for counting using domain disentanglement. We demonstrate the improved cell counting accuracy by predicting density maps using a small fraction of auto-generated exemplars on datasets with only weak dot annotations for cells.

Publications

1. **Veena Dodbballapur**, Rajanish Calisa, Yang Song, and Weidong Cai, "Automatic Dropout for Deep Neural Networks," Neural Information Processing: 27th International Conference, ICONIP 2020, (pp. 185-196), Springer, Cham, doi: 10.1007/978-3-030-63836-8_16.
Received best paper finalist award.
2. **Veena Dodbballapur**, Yang Song, Heng Huang, Mei Chen, Wojciech Chrzanowski, and Weidong Cai, "Mask-Driven Mitosis Detection in Histopathology Images," IEEE 16th International Symposium on Biomedical Imaging, ISBI 2019, (pp. 1855-1859), doi: 10.1109/ISBI.2019.8759164.
3. **Veena Dodbballapur**, Yang Song, Heng Huang, Mei Chen, Wojciech Chrzanowski, and Weidong Cai, "Dual-Stage Domain Adaptive Mitosis Detection for Histopathology Images," Digital Image Computing: Techniques and Applications, DICTA 2020, (pp. 1-7), doi: 10.1109/DICTA51227.2020.9363411.
4. **Veena Dodbballapur**, Dongnan Liu, Yang Song, and Weidong Cai, "Cross-Domain Exemplars for Counting," IEEE International Symposium on Biomedical Imaging, ISBI 2024, (Accepted for presentation at the 21st International Symposium on Biomedical Imaging, ISBI 2024).
5. **Veena Dodbballapur**, Dongnan Liu, Rajanish Calisa, Yang Song, and Weidong Cai, "Spatio-Temporal Region-Specific Features for Person Re-Identification," IEEE International Conference on Multimedia and Expo, ICME 2024, (Under Review).

Authorship Attribution Statement

1. Chapter 2 of this thesis is published as (Dodballapur et al., 2020a) .

I designed this study, analysed the data, and wrote the drafts with the co-authors.

2. Chapter 3 of this thesis is published as (Dodballapur et al., 2019).

I designed this study, analysed the data, and wrote the drafts with the co-authors.

3. Chapter 4 of this thesis is published as (Dodballapur et al., 2020b).

I designed this study, analysed the data, and wrote the drafts with the co-authors.

4. Chapter 5 of this thesis has been accepted at the IEEE International Symposium on Biomedical Imaging (ISBI 2024) and is not available online.

I designed this study, analysed the data, and wrote the drafts with the co-authors.

5. Chapter 6 of this thesis is currently under review and has been submitted to IEEE International Conference on Multimedia and Expo (ICME 2024) and is not available online.

I designed this study, analysed the data, and wrote the drafts with the co-authors.

Student name: Veena Dodballapur

Date: 10/01/2024

As supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Supervisor name: Weidong Cai

Date: 10/01/2024

Acknowledgements

I would like to express my deepest gratitude and appreciation to all the people who helped and supported me in my PhD journey.

First and foremost, I am extremely grateful to my supervisor, Associate Professor Weidong (Tom) Cai for his continuous encouragement and invaluable guidance throughout my PhD. Especially during COVID, he took special care, supported and motivated me in my research work. Without his support, this endeavour would not have been possible.

Further, I am grateful to my associate supervisors Dr. Dongnan Liu and Associate Professor Yang Song (University of New South Wales) for providing technical insight and research guidance. Our meetings and conversations on a regular basis refined my research direction and also helped me to think outside the box. I really appreciate their guidance.

I would like to extend my appreciation, for the support I received from my research collaborators: Professor Heng Huang (University of Pittsburgh), Dr. Mei Chen (Microsoft Cloud & AI), Professor Wojciech Chrzanowski (University of Sydney), and Rajanish Calisa (University of Technology Sydney).

I am very grateful to Dr. Jeroen Vendrig, Danny Cohen, and other colleagues at ProofTec for providing the required flexibility and generous support to complete my research work. Special mention and thanks to my friend Dr. Jeroen Vendrig for his emotional support and encouragement throughout my study.

My deepest gratitude to my parents, husband and daughter for their unconditional love, support, and motivation. I am blessed, to have my parents provide me with the right values and education, so that I could pursue my doctoral studies. I am deeply indebted to my husband, who helped and stood by me during this entire PhD journey. Without my family's support, this journey would not have been possible. Finally, I would like to dedicate this thesis to God and my family.

CONTENTS

Student Plagiarism: Compliance Statement	ii
Abstract	iii
Publications	vi
Authorship Attribution Statement	vii
Acknowledgements	ix
Contents	x
List of Figures	xii
List of Tables	xx
Chapter 1 Introduction	1
1.1 Supervised Learning	2
1.2 Data-Efficient Deep Learning	10
1.3 Applications	18
1.4 Evaluation Metrics	28
1.5 Challenges	33
1.6 Thesis Contributions and Organization	36
Part 1. Advanced Learning-based Fully Supervised Methods	40
Chapter 2 Automatic Dropout for Deep Neural Networks	41
2.1 Introduction	41
2.2 Method	46
2.3 Experiments and Results	52
2.4 Summary	59
Chapter 3 Mask-Driven Mitosis Detection in Histopathology Images	60

3.1	Introduction	61
3.2	Method	65
3.3	Experiments and Results	70
3.4	Summary	77
Part 2. Advanced Learning-based Weakly Supervised Methods		78
Chapter 4 Dual-Stage Domain Adaptive Mitosis Detection for Histopathology		
	Images	79
4.1	Introduction	80
4.2	Method	84
4.3	Experiments and Results	89
4.4	Summary	97
Chapter 5 Cross Domain Exemplars for Cell Counting		98
5.1	Introduction	99
5.2	Method	104
5.3	Experiments and Results	111
5.4	Summary	116
Chapter 6 Spatio-Temporal Region-Specific Features for Person Re-Identification		117
6.1	Introduction	118
6.2	Method	121
6.3	Experiments and Results	128
6.4	Summary	141
Part 3. Conclusions and Future Work		142
Chapter 7 Conclusions and Future Work		143
7.1	Conclusions	143
7.2	Future Work	145
Bibliography		148

List of Figures

- 1.1 Visual recognition tasks in computer vision: (a) "Image Classification" involves assigning categorical class labels to an image; (b) "Object Detection" not only predicts categorical labels but also localizes each object instance via bounding boxes; (c) "Semantic Segmentation" aims to predict categorical labels for each pixel, without distinguishing between object instances; (d) "Instance Segmentation" represents a specialized form of object detection that distinguishes different object instances through pixel-level segmentation masks (Wu et al., 2020). 2
- 1.2 Popular approaches to image segmentation: (a) Fully convolutional networks (FCN) applied for pixel-wise semantic segmentation (Long et al., 2015a); (b) SegNet architecture based on VGG-16, followed by the deconvolution network (Badrinarayanan et al., 2017); (c) U-Net architecture which is able to achieve precise segmentation with fewer training samples (Ronneberger et al., 2015); (d) The Segmentor predicts class predictions per pixel and the adversarial network takes a label map as input, and outputs a label, which predicts whether the image is ground truth or synthetic (Luc et al., 2016). 4
- 1.3 A neural network with standard dropout (Srivastava et al., 2014). (Left): A standard neural network with two hidden layers, one input and output layer. Note the connections between neurons of the two layers. Every neuron in a layer is connected to every other neuron in the layer below and above it. (Right): The effect of dropout is the dropping out of one or more neurons in the hidden layers, and as a result their connections are also dropped. The dropped neurons are shown as dotted circles with a cross in it. 7
- 1.4 Modern augmentation strategies (Yun et al., 2019): (a) Original image; (b) Mixup augmentation; (c) Cutout or random-erase augmentation; (d) CutMix augmentation. 10
- 1.5 A generic pseudo-labeling method, where a network is trained with both labeled and unlabeled data and pseudo-labels are assigned to the unlabeled data, based on

- learned feature representations. The pseudo-labels are then used for future training phases for fully supervised training (Yang et al., 2022). 14
- 1.6 The proposed architecture includes a deep feature extractor (green) and a deep label predictor (blue), which together form a standard feed-forward architecture. Unsupervised domain adaptation is achieved by adding a domain classifier (red) connected to the feature extractor via a gradient reversal layer that multiplies the gradient by a certain negative constant during the backpropagation-based training (Ganin et al., 2016). 15
- 1.7 (left) The state-of-the-art dual stage Mask R-CNN network which is based on Faster R-CNN (He et al., 2017), (right) single-stage monolithic networks YOLO and SSD (Liu et al., 2016b). 18
- 1.8 Density-map based Crowd Scene Counting Method (Zhang et al., 2015). In this method patches on images containing people, are used to train a CNN with two training objectives: (1) counting the crowd in the patch; (2) generating a density map for the patch. The multi-task training proposed here improves the overall accuracy of the crowd counting. 20
- 1.9 Person ReID representation learning from images captured from multiple cameras and matching a query image of a specific person with a gallery (Ye et al., 2021). 22
- 1.10 Different phases of Mitosis Cells (left to right): Prophase, Metaphase, Anaphase, and Telophase (Beevi et al., 2019). Note that the shape of a nucleus varies as the cell goes through the 4 stages, and in the last phase where there are two divided nuclei, the count of cells is still one. These morphological variations in the cell nucleus poses a great challenge for automatic cell counting. 24
- 1.11 Mitosis Cells and its mimics (Li et al., 2018a): (a) The first three rows, are mitosis cells and the third row are cells undergoing telophase and have two nuclei; (b) Mimics of mitosis cells are in the fourth row. They are similar to mitosis cells in appearance. 26
- 1.12 Breast tissue samples stained with hematoxylin and eosin from the Tumor Proliferation Assessment Challenge (TUPAC) dataset (Tellez et al., 2018). The images come from different laboratories and as a result have different staining. 27

- 2.1 Automatic dropout rate determination at the three layers of a neural network. The distribution of neuron activations is approximated as a Gaussian with the determined standard deviation. The dropout rate for each layer is determined from this Gaussian distribution. The dropped neurons are shown with a red cross against them. The standard deviation of activations at layer $N - 1$ is determined as 0.9, and for layer N , the dropout rate is sampled from a Gaussian distribution with $SD = 0.9$. Similarly, for layer $N + 1$, the dropout rate is sampled from Gaussian with $SD = 0.4$. 45
- 2.2 Adaptive dropout based on clustering working in combination with automatic dropout rate selection. The clustering of activations allows different dropout rates to be applied selectively to activations within a cluster, based on the strength of the activation. At each layer, as illustrated, three clusters are determined, based on the neuron activations. Different dropout rates are applied to each cluster, based on the strength of the activation of each cluster. For determining the strength of activation, the cluster centroid's activation is used. The adaptive dropout rate applied is derived from an empirically determined fixed dropout rate or from a distribution of activations of the previous layer. As an example, for layer N , there are three clusters with centroids 0.15, 0.2 and 0.3, and the dropout rate of 0.3 is adaptively applied to each cluster. This rate is sampled from a Gaussian of $SD = 0.6$. 51
- 2.3 Sample images from (Left) CIFAR-10 dataset (Krizhevsky et al., 2009) and (Right) Fashion-MNIST dataset (Xiao et al., 2017). CIFAR-10 dataset has 10 random images per row for each of the 10 classes in the dataset. Fashion-MNIST dataset has 3 rows of images per class and has 10 classes in the dataset. 58
- 3.1 A three-stage process of mitosis detection. In the first stage, RGB images are transformed to other color spaces and specific color channels are selected using histogram analysis. In the second stage, a blue ratio image is determined which is useful in differentiating nuclei from background. In the third stage, morphological and statistical features are computed to train a classification model for mitosis classification (H.Irshad, 2013). 62
- 3.2 Mask R-CNN is a flexible and general framework for object detection and instance segmentation. It belongs to the family of region-proposal networks and an extension of Faster R-CNN. Objects in an image are detected, while simultaneously

- predicting high-quality segmentation mask at pixel-level, for each detected object. This instance segmentation achieves better localisation (He et al., 2017). Our two-stage mitosis detection framework, uses Mask R-CNN in the first stage to predict masks for the datasets. 63
- 3.3 DeepMitosis uses a 2-step mitosis detection approach (Li et al., 2018a). A deep segmentation model based on FCN is used to generate candidate bounding boxes from weakly annotated datasets. Then a Faster R-CNN network is used in the deep detection model to predict bounding boxes, followed by deep verification step, which performs mitosis classification. 65
- 3.4 Illustration of mask generation using Mask R-CNN: (a) The image shows the High Power Field image with ground truth annotation in yellow; (b) The image shows masked mitosis cells in green, predicted by Mask R-CNN. The manually marked bounding boxes in yellow show the mitosis cells matching the ground truth mitosis cells. Other non-marked cells are hard negatives mitosis cells which are false positives. 67
- 3.5 Xception network (Chollet, 2017) which is used for the mitosis detection stage. Xception network has depth-wise separable convolution layers. The network has three distinct parts, entry flow, middle flow and exit flow. The first two are repeated eight times. Hence, the training data goes through eight sets of entry and middle flow parts before it is passed into the single exit flow part. For the mitosis detection stage, we use a modified version of this network where the last two fully connected layers and the classification layer are removed. 69
- 3.6 Mitosis detection stage where candidate detections from stage 1 are classified as mitotic or non-mitotic. Hand-crafted features extracted from the masks generated from the Mask R-CNN network are combined with deep features to classify the candidate cells. The input images are RGB image and the binary mask image. The learned deep features are extracted using the Xception network or the VGG-16 network. The combination of learned deep features and hand-crafted features are used to train a network with a fully connected layer followed by a mitosis classification layer. 70

- 3.7 Sample image patches from the Mitosis 2012 and 2014 dataset showing mitosis cells. The first two rows correspond to Mitosis 2012 dataset and the last two rows are from the Mitosis 2014 dataset. Structural similarity can be noticed in the cells from the two datasets, but the overall appearance differs due to non-standard staining and tissue preparation process. 72
- 3.8 Masks generated for sample image patches from the weakly annotated ICPR 2014 Mitosis dataset by our Mask R-CNN model trained on ICPR 2012 dataset. The first row contains the image patches, and the second row has the corresponding masks applied to the image patches to isolate the cells. Note that the masks generated are accurate in all cases except the fifth image from the left, where the cell is in the telophase stage. 74
- 4.1 Effect of non-standard tissue preparation and staining. (Left), a whole slide image from ICPR 2012 dataset and (right) a whole slide image from ICPR 2014 dataset. In both instances, examples of mitosis and non-mitosis cells are shown using yellow bounding boxes. 80
- 4.2 Domain Adaptive Mask Generation: the last layer of the ResNet-50 backbone connects both the Mask R-CNN head as well as the domain adaptation branch. A gradient reversal layer has been introduced to the Mask R-CNN network to enable the network to learn domain invariant features. The domain adaptive Mask R-CNN is used in the first stage of our dual-stage domain adaptive mitosis detection framework. 87
- 4.3 Domain Adaptive Mitosis Detection Stage: last layer of the VGG-19 network connects to both the mitosis classification branch as well as the domain adaptation branch. This forms the second stage of our dual-stage domain adaptive mitosis detection framework. As in the first stage the gradient reversal layer is used to let the VGG-19 network learn domain invariant features. 88
- 4.4 Sample whole slide images from ICPR 2012 and 2014 datasets: (a) ICPR 2012 using Aperio scanner; and (b) ICPR 2012 dataset using Hamamatsu scanner, (c) ICPR 2014 dataset using Aperio scanner. 90
- 4.5 Examples of masks predicted by the domain adaptive Mask R-CNN (DA) compared against ground-truth images from the ICPR 2012 dataset (scanner H).

- Visually, it can be seen that masks generated by domain adaptive Mask R-CNN closely resemble ground-truth in both shape and size when compared to non-domain adaptive Mask R-CNN (non-DA). 93
- 4.6 Examples of the domain adaptive mitosis detection stage (second stage) results on the ICPR 2012 dataset (scanner H) by both domain adaptive (DA) and non-domain adaptive (non-DA) models. When predictions are compared against the ground-truth, the DA model has more correct predictions than the non-DA model. 95
- 5.1 Cross-domain exemplar generation and selection framework: (a) Two Faster R-CNN networks are trained with RGB, and background subtracted images, respectively. The detections which are bounding boxes predicted by the Faster R-CNN network are merged based on overlap and confidence scores; (b) Candidate bounding boxes from (a) are clustered based on features extracted from a ResNet-50 backbone and exemplars are selected from cluster centroids based on their respective confidence scores. 106
- 5.2 Illustrative exemplars selected from images of three different datasets we used in our evaluation: (a) DCC dataset, (b) MBM dataset; and (c) VGG dataset. For each dataset, we illustrate with bounding boxes, the exemplars that have been chosen. Note that for the VGG dataset, for clarity, we have shown the cells in red colour instead of blue. 107
- 5.3 Result of background subtraction step applied to cell images from the three datasets we used for our evaluation. The first row contains the original RGB images, and the second row has the corresponding background subtracted images: (a) VGG dataset; (b) DCC dataset; (c) MBM dataset. 108
- 5.4 Results of K -means clustering of candidate exemplars to account for variation in cell appearance. (a) Image with candidate exemplars determined by merging bounding box predictions from the two Faster R-CNN networks. (b) Three clusters X, Y, Z as determined by K -means using ResNet-50 features extracted from image patches of the candidate exemplars. The clusters are representative of the variations in the cells. 109
- 5.5 Results of determining candidate detections through confidence-based merging: (a) Detections obtained from bounding box predictions on the background subtracted

- image; (b) Detections obtained from bounding box predictions on the RGB image; (c) Merged candidates, combining detections from RGB, and background subtracted images. 110
- 5.6 Datasets used in the experiments: (a) MBM Dataset; (b) VGG Synthetic Dataset; (c) DCC Dataset. 112
- 6.1 Architecture of our ReID learning framework which uses region-specific augmentation to train the network. The implementation of hybrid memory is illustrated to store region-specific features which are accumulated both temporally and spatially. The temporal accumulation creates intra-region features and spatial accumulation creates inter-region features. The cluster centers and pseudo-labels generated from ReID features learnt from region-specific augmentation are stored in hybrid memory. 123
- 6.2 An example image with the three identified regions: (1) the original RGB image, and region-specific augmentation is shown in (2) head region is augmented; (3) torso region is augmented and (4) leg region is augmented. 124
- 6.3 Region-specific features combined together to generate intra-region features. Two different augmentations of the head region are shown and the corresponding ReID features (f_1, f_2) are combined to create intra-region features (f_{12}). 124
- 6.4 Region-specific features combined to produce an inter-region feature. Intra-region features which have been temporally accumulated in hybrid memory for head region (f_1), torso region (f_2) and leg region (f_3) are combined spatially to create an inter-region feature (f_{123}). 125
- 6.5 Outline of how region-specific features are accumulated in hybrid memory. In Step 1, intra-region features are accumulated during training, and in Step 2, the accumulated intra-region features are combined to form inter-region features. 126
- 6.6 Implementation of feature accumulation in one experimental setting at the epoch level during training. Intra-region features are accumulated within the epoch. Inter-region features are obtained at the end of the epoch by combining intra-region features. 127
- 6.7 Sample images from the four datasets we have used in our evaluation: (a) Market-1501 (Zheng et al., 2015) dataset; (b) DukeMTMC-reID (Ristani et al.,

- 2016) dataset; (c) MSMT17 (Wei et al., 2018) dataset showing variations in lighting, scene, pose and occlusions; (d) PersonX (Sun and Zheng, 2019) dataset with some sample images in various background and pose and the 3 colour and scene variations. 130
- 6.8 Region-specific features reduce noise as illustrated and hence result in robust clusters and pseudo-labels with less noise. 134
- 6.9 Ablation experiments: (a) A region is chosen randomly with uniform probability at the beginning of the epoch and used for region-specific augmentation for all images within that epoch; (b) A region is chosen in a sequential manner at the beginning of the epoch and applied for all images within the epoch to learn region-specific features; (c) A region is chosen randomly for each image within the epoch, thus creating different region-specific augmentations for individual images; (d) A constant region-specific augmentation is performed and in our case region 2 is chosen which is the torso. 140

List of Tables

1.1	Confusion Matrix.	28
2.1	Results comparing our method with the state-of-the-art method jumpout, and the standard dropout, for different variants of Wide ResNet ($width = 18, depth = 8$ and ($width = 28, depth = 10$) and popular datasets. CIF10, CIF100 and FASHN refer to the CIFAR-10, CIFAR-100, Fashion-MNIST datasets, respectively. The result for M_1 outperforms the state-of-the-art for both Wide ResNet variants, and $M_1 + M_2$ outperforms M_1 and state-of-the-art.	55
2.2	Results our method compared with standard dropout for the v3 network and popular datasets. CIF10, CIF100 and FASHN refer to the CIFAR-10, CIFAR-100, Fashion-MNIST datasets, respectively. The result for M_1 outperforms the standard dropout, and $M_1 + M_2$ outperforms M_1 and the standard dropout.	55
2.3	Results of the ablation experiments comparing our method in different settings with jumpout and the standard dropout, using Wide ResNet variants ($width = 18, depth = 8$ and ($width = 28, depth = 10$) and popular datasets. CIF10, CIF100 and FASHN refer to the CIFAR-10, CIFAR-100, Fashion-MNIST datasets, respectively. The result for $M_1 + M_2$ or M_2 alone outperforms all other methods including state-of-the-art. Method M_2^{min} performs worse than M_2 .	57
2.4	Results of the ablation experiments on Wide ResNet:40_4 ($width = 40, depth = 4$) using popular datasets, comparing our method in different settings, and comparing it with jumpout and standard dropout. CIF10, CIF100 and FASHN refer to the CIFAR-10, CIFAR-100, Fashion-MNIST datasets, respectively. The result for $M_1 + M_2$ outperforms the all other methods including state-of-the-art. Method M_2^{min} performs worse than M_2 .	57
3.1	Results of our two-stage mitosis detection framework in two different settings and comparison with state-of-the-art methods on the ICPR 2012 dataset. Our method, with	

- the Xception network in the second stage, achieves the best precision and F1 score, and with the VGG-16 network, achieves the equal-best recall score. 73
- 3.2 Results of our two-stage mitosis detection framework in two different experimental settings and comparison with state-of-the-art methods on the ICPR 2014 dataset where evaluation was performed on the ICPR 2014 validation set. Our method based on VGG-16 outperforms state-of-the-art DeepMitosis and RCNN in both recall and F1 scores. Our implementation, based on Xception, achieved the second-best recall score. 75
- 3.3 Results of ablation experiments of our two-stage mitosis detection framework comparison with state-of-the-art methods on the ICPR 2012 dataset. In particular, when we used learned deep features only to train our mitosis detection stage, our method matched DeepMitosis in F1 score and showed marginal improvement in precision score. When using deep learned features, the performance was only marginally lower when compared to our best result which used both learned deep features and hand-crafted features (first two rows). 76
- 3.4 Results of the ablation experiments of our two-stage mitosis detection framework, and comparison with state-of-the-art methods, on the ICPR 2012 dataset. In particular, we compared the results of using a simple network with 5 convolutional layers, with VGG-16 and Xception network. Across all metrics, the results were lower. 77
- 4.1 Results of the dual-stage domain adaptive mitosis detection on ICPR 2012 Scanner H test set comparing the domain adaptive and non-domain adaptive settings. In the domain adaptive setting, the framework was trained using ICPR 2012 scanner A images with supervision, and ICPR 2012 scanner H training images with no supervision. ICPR 2012 (A) refers to scanner A images and ICPR 2012 (A, H) refers to both scanner A and H images. The results show that the domain adaptive model out-performs the non-domain adaptive model. 92
- 4.2 Results of the dual-stage domain adaptive mitosis detection on the ICPR 2014 validation split (scanner A) and comparison with the state-of-the-art. In the domain adaptive setting, the framework was trained using ICPR 2012 scanner A images with supervision and ICPR 2014 scanner A training images with no supervision. ICPR 2012 (A) refers to scanner A images of ICPR 2012 dataset, and ICPR 2014 (A) refers to scanner A images of ICPR 2014 dataset. In both precision and F1 scores, our domain

adaptive method outperforms state-of-the-art even though the other methods have been trained in a supervised setting for ICPR 2014 dataset.	94
4.3 Results of the dual-stage domain adaptive mitosis detection on the ICPR 2012 (scanner H) test set, and comparison with the state-of-the-art. We note that the recall rate of the domain adaptive model gets closer to state-of-the-art which has been trained with full supervision on the ICPR 2012 (scanner H) training dataset.	96
5.1 Results of cross-domain automatic cell counting in different settings. For each experiment, we note the target and source datasets, with the source indicated in brackets. E.g. DCC(VGG) indicates that DCC is the target and VGG is the source dataset. "Original" indicates cross-domain training done with RGB microscopy images, "Pre-processed" with background subtracted images and "merged" which uses a combination of original RGB and background-subtracted images.	114
5.2 Results of cross-domain automatic cell counting compared with state-of-the-art methods. The "Cells" column indicates the approximate number of cells used at training time which are human annotated (in state-of-the-art). For our method, the cells column indicates the number of suggested exemplars used for density map generation.	115
6.1 Results of our experiments with Duke as the source dataset and Market-1501 as the target dataset. We report mAP, RANK-1 (R1), RANK-5 (R5), RANK-10 (R10) results for our method, in comparison with state-of-the-art methods. The last row provides our implementation results. The best result is shown in bold, and the second-best result is underlined.	135
6.2 Results of our experiments with Market-1501 as the source dataset and Duke as the target dataset. We report mAP, RANK-1 (R1), RANK-5 (R5), RANK-10 (R10) results for our method, in comparison with state-of-the-art methods. The last row provides our implementation results. The best result is shown in bold, and the second-best result is underlined.	136
6.3 Results of our experiments with PersonX as the source dataset and Market-1501 as the target dataset. We report mAP, RANK-1 (R1), RANK-5 (R5), RANK-10 (R10) results for our method, in comparison with state-of-the-art methods. The last row provides our implementation results. The best result is shown in bold, and the second-best result is underlined.	136

- 6.4 Results of our experiments with Market-1501 as the source dataset and MSMT as the target dataset. We report mAP, RANK-1 (R1), RANK-5 (R5), RANK-10 (R10) results for our method, in comparison with state-of-the-art methods. The last row provides our implementation results. The best result is shown in bold, and the second-best result is underlined. 137
- 6.5 Results of our experiments with Duke as the source dataset and MSMT as the target dataset. We report mAP, RANK-1 (R1), RANK-5 (R5), RANK-10 (R10) results for our method, in comparison with state-of-the-art methods. The last row provides our implementation results. The best result is shown in bold, and the second-best result is underlined. 138
- 6.6 Results of mAP, RANK-1 (R1), RANK-5 (R5), and RANK-10 (R10) in the Market-1501 to Duke branch of the ablation experiment. The second row (ResNet-50) outlines the results when the backbone is ResNet-50. The third row (ResNet-50-Part) outlines the results when only one large region was used, and ResNet-50 was the backbone. The best results (ResNet50-IBN-a), shown in bold, are the first set of results with ResNet50-IBN-a as a backbone. 138
- 6.7 Results of mAP, RANK-1 (R1), RANK-5 (R5), and RANK-10 (R10) in the Duke to Market-1501 branch of the ablation experiment. The second row (ResNet-50) outlines the results when the backbone is ResNet-50. The third row (ResNet-50-Part) outlines the results when only one large region was used, and ResNet-50 was the backbone. The best results (ResNet50-IBN-a), shown in bold, are the first set of results with ResNet50-IBN-a as a backbone. 139
- 6.8 Results of mAP, RANK-1 (R1), RANK-5 (R5), and RANK-10 (R10) in the Market-1501 to Duke branch performed on the MMT network. The first row has the results for two-stage feature accumulation method as part of the main experiment. The second row is baseline result for the MMT ReID network. Rows 3-5 are of the variations of our method as described in the Ablation Experiments section. 141

CHAPTER 1

Introduction

Computer vision encompasses visual content analysis tasks, to extract and interpret visual information from images and videos of the real world. It involves algorithms, techniques, and systems, which enable computers to develop an understanding of the world, from digital images. Over the years, technology for computer vision applications, such as classification, segmentation, object detection, and recognition, have advanced from using hand-crafted features, to machine-learned features. The early algorithms relied on sophisticated methods to obtain the most appropriate hand-crafted features for visual content analysis tasks, such that classification, segmentation, and object recognition could be accomplished accurately.

Techniques, such as Scale Invariant Feature transform (SIFT) (Lowe, 2004), its variants such as PCA-SIFT (Ke and Sukthankar, 2004), Histogram of Gradients (HOG) (Dalal and Triggs, 2005), colour features, and gradient features, have been applied successfully for a variety of tasks, ranging from, classification, face recognition, and image segmentation, to medical image analysis used in cancer detection and cell and tumour classification. Early learning-based techniques, such as the Viola Jones (Viola and Jones, 2001) detector, have been applied for object detection and face detection tasks. The Viola Jones detector uses integral images, Haar-like features, and cascade classifiers to solve detection tasks. Histogram of Gradients (HOG) (Dalal and Triggs, 2005), which detects object shapes at various scales, has been successfully applied to pedestrian detection, in conjunction with Support Vector Machines (SVMs) (Sotelo et al., 2006), (Papageorgiou and Poggio, 2000).

However, in the last decade, deep learned features using neural networks including convolutional neural networks (CNNs), have achieved state-of-the-art performance on a variety of visual content analysis tasks, and have overtaken hand-crafted features and early machine learning methods. A deep neural network (DNN) has multiple layers of interconnected nodes or neurons, arranged in a hierarchical fashion. The term, "deep", indicates the existence of two

or more hidden layers between the input and output layers. A convolutional neural network (CNN) is a type of DNN which is designed to work with image data to perform visual content analysis tasks. Deep learning techniques learn efficient and complex representations directly from image data (pixels), which can then be applied to a variety of visual analysis tasks, including classification, detection, segmentation, and object counting.

In the following sections, we provide an overview of various deep learning algorithms, data supervision methods and their applications to solving computer vision tasks, as relevant to this thesis. We discuss supervised learning and its applications, weakly supervised learning, and unsupervised learning with domain adaptation. We will also explore their respective challenges in training robust deep learning models and model generalization.

1.1 Supervised Learning

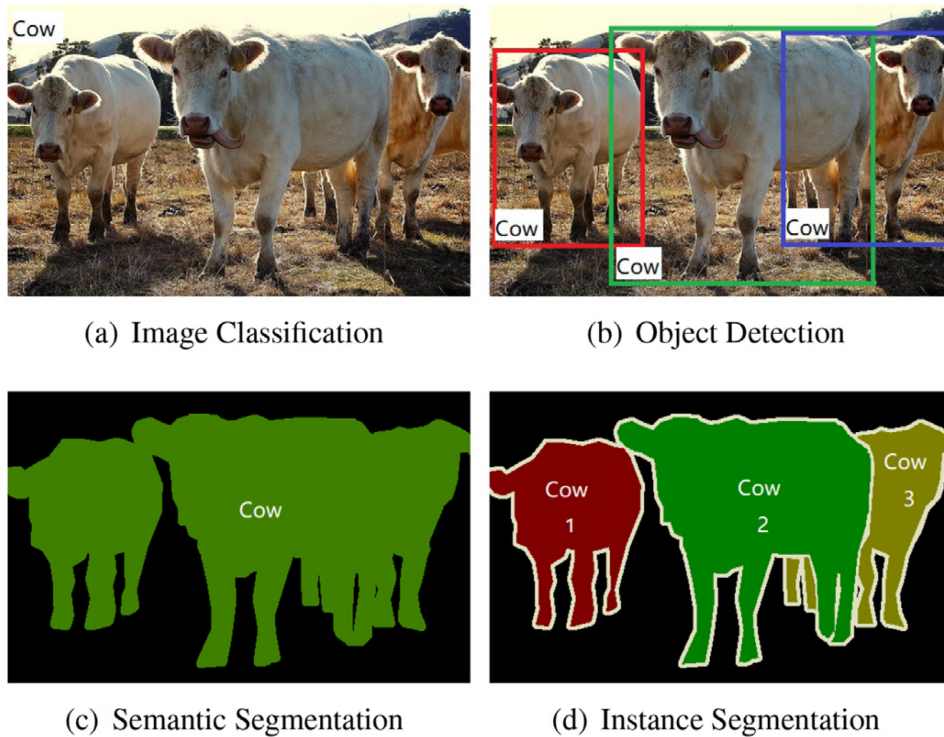


FIGURE 1.1: Visual recognition tasks in computer vision: (a) "Image Classification" involves assigning categorical class labels to an image; (b) "Object Detection" not only predicts categorical labels but also localizes each object instance via bounding boxes; (c) "Semantic Segmentation" aims to predict categorical labels for each pixel, without distinguishing between object instances; (d) "Instance Segmentation" represents a specialized form of object detection that distinguishes different object instances through pixel-level segmentation masks (Wu et al., 2020).

One of the most key areas of focus in machine learning research, is the extraction of useful information from past experience. Past information is represented as training data, and is used to train a machine learning model, which is then used to predict outcomes from unseen data. The branch of machine learning which uses this technique is called supervised machine learning. Supervised deep learning techniques require a large amount of human-annotated training data, to learn a generalized model.

Supervised learning has been applied to visual content analysis tasks, such as image classification (Russakovsky et al., 2015), object detection, and recognition (Zhang et al., 2013), (Everingham et al., 2010), as well as semantic segmentation (Russakovsky et al., 2015). Specifically, image classification focuses on recognizing the semantic categories of objects within an image as shown in Figure 1.1 (a). Object localization involves identifying the location of the objects within an image using a bounding box, or a polygon. Object detection goes a step further, by predicting both object categories, as well as the precise location of each object using bounding boxes, as shown in Figure 1.1 (b). Semantic segmentation as shown in Figure 1.1 (c), on the other hand, predicts classifiers on a pixel-by-pixel basis, assigning specific category labels to each pixel, and thereby offering a more comprehensive and fine grained classification of an image. However, unlike object detection, semantic segmentation does not differentiate between multiple objects of the same category. "Instance segmentation" as shown in Figure 1.1 (d), is an emerging technique, combining object detection and semantic segmentation, to distinguish objects and allocate each of them, a distinct, categorical pixel-level mask. Thus, instance segmentation is a specialized form of object detection, emphasising pixel-level localization, rather than bounding-box localization.

As discussed above, supervised learning has thus been applied successfully to several visual content analysis tasks. However, there continues to be many challenges in the field of visual recognition. Examples of challenges include requirement of large, annotated datasets, pixel-level annotations for segmentation tasks, and model generalization.

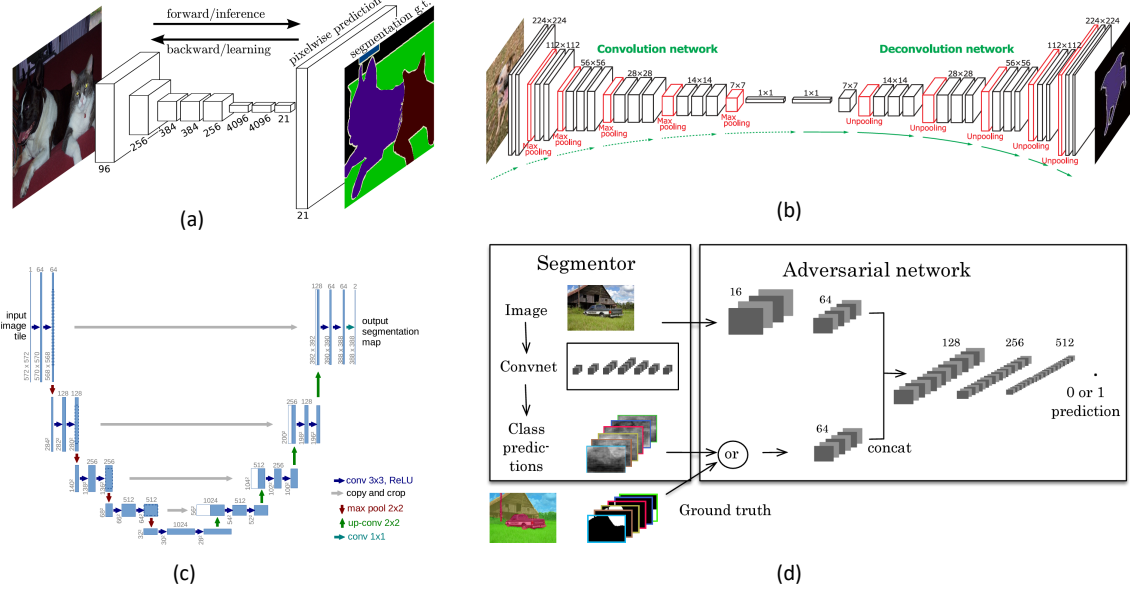


FIGURE 1.2: Popular approaches to image segmentation: (a) Fully convolutional networks (FCN) applied for pixel-wise semantic segmentation (Long et al., 2015a); (b) SegNet architecture based on VGG-16, followed by the deconvolution network (Badrinarayanan et al., 2017); (c) U-Net architecture which is able to achieve precise segmentation with fewer training samples (Ronneberger et al., 2015); (d) The Segmentor predicts class predictions per pixel and the adversarial network takes a label map as input, and outputs a label, which predicts whether the image is ground truth or synthetic (Luc et al., 2016).

1.1.1 Semantic Segmentation

Image segmentation is a key field of study within supervised learning, with wide-ranging applicability to visual recognition tasks. Image segmentation can be formulated as the problem of classifying pixels with semantic labels (semantic segmentation) or partitioning of individual objects (instance segmentation). Semantic segmentation performs pixel-level labelling with a set of object categories (e.g., human, car, tree, sky) for all image pixels; thus, it is generally a more demanding task when compared to whole-image classification, which predicts a single label for the entire image.

The motivation for semantic segmentation comes from the need for visual understanding systems, where images are partitioned into multiple regions and objects (Szeliski, 2022). It plays a central role in a broad range of vision-based applications (Forsyth and Ponce, 2002), including medical image analysis, such as tumour cell localization and tissue measurement, as

well as autonomous vehicles and video surveillance. Instance segmentation extends the initial idea of semantic segmentation, by detecting each object of interest in the image, separately (e.g. individuals). Numerous image segmentation algorithms have been developed in the literature, from the earliest methods, such as thresholding (Otsu, 1979), histogram-based bundling, region growing (Nock and Nielsen, 2004), K -means clustering (Dhanachandra et al., 2015), Watershed methods (Najman and Schmitt, 1994), to more advanced algorithms, such as active contours (Kass et al., 1988), graph cuts (Boykov et al., 2001), Conditional and Markov random fields (Plath et al., 2009), and Sparsity-based (Starck et al., 2005) methods.

More recently, deep learning models have led to advancement in image segmentation, with CNN-based models achieving the highest accuracy rates on popular benchmarks. Several CNN architectures have been proposed for performing image segmentation. "Fully Convolutional Network" (FCN) (Long et al., 2015a), a deep learning approach, depicted in Figure 1.2 (a), is an architecture based on convolutional layers only. FCN predicts a segmentation image, whose size is the same as that of the input image and has been applied successfully to a variety of segmentation problems (Wang et al., 2018), (Yuan et al., 2017), (Liu et al., 2016a), (Li et al., 2017c). Semantic segmentation using a deconvolution network consisting of deconvolution and un-pooling layers (Noh et al., 2015) has been proposed to better handle objects at different scales. A similar approach, namely SegNet (Badrinarayanan et al., 2017), depicted in Figure 1.2 (b), uses an encoder-decoder architecture. SegNet is similar to a deconvolution network, in that the core, trainable segmentation part of SegNet, is an encoder, structurally identical to the 13 convolutional layers of the VGG16 network. The corresponding decoder network is followed by a pixel-wise classification layer.

"U-Net" (Ronneberger et al., 2015), shown in Figure 1.2 (c), is a segmentation architecture, comprising of two parts, a contracting path to capture context, and a symmetric expanding path that enables precise localization. U-Net has been shown to be very efficient in segmenting biological microscopy images. Multi-scale image analysis is a well-known technique within computer vision and has made its way into various neural network architectures, such as the Feature Pyramid Network (FPN) (Lin et al., 2017). More recently, Generative Adversarial Networks (GANs) have been applied to a wide range of tasks in computer vision, including image segmentation. The study (Luc et al., 2016), proposes an adversarial training approach for semantic segmentation, depicted in Figure 1.2 (d), in which they trained a convolutional

semantic segmentation network, along with an adversarial network, that discriminates between ground-truth segmentation maps, and those generated by the segmentation network. Figure 1.2 illustrates the most important and representative architectures used for semantic segmentation as described in this section.

1.1.2 Object Detection

Object detection and recognition, as an important field of study in computer vision, has naturally emerged from object classification and image segmentation approaches. The task of generic object detection (Zhang et al., 2013), can be defined as follows: when presented with an image, the objective is to determine whether there are occurrences of objects belonging to pre-defined categories. If any of the instances are identified, the aim is then to provide spatial location and size of the detected object. Hence, the broader problem of object detection involves identifying and localizing all the objects present in an image, encompassing the challenges of both image segmentation and classification (Everingham et al., 2010), (Russakovsky et al., 2015), (Andreopoulos and Tsotsos, 2013).

Object detection research is primarily focused on highly structured objects with definite shape and texture characteristics, such as cars, faces, bicycles, aircraft, humans, and animals, as opposed to unstructured objects, such as the sky, grass, and clouds. The spatial location and size of an object can be roughly defined using a bounding box (this is an axis-aligned rectangle that tightly encloses the object) (Russakovsky et al., 2015), (Everingham et al., 2010), a precise pixel-wise segmentation mask (Zhang et al., 2013), or a closed polygon defining the object boundary (Lin et al., 2014), (Russell et al., 2008). Currently, the use of bounding boxes is predominant in all object detection algorithms ((Russakovsky et al., 2015), (Everingham et al., 2010), (Opelt et al., 2006)). However, as recent advances have been made in pixel-wise object segmentation, we anticipate that future challenges will lie in improving the accuracy of pixel-level classification (Lin et al., 2014). There are several challenges associated with generic object detection. While the goal of object classification, shown in Figure 1.1 (a), is to determine the presence of an object within an image, there is no need for precise localization. However, in the case of object detection, precise localization makes it a more challenging task when compared with classification. To achieve good accuracy, pixel-wise annotations become important while training object detection models. Figure 1.1 (c), which depicts generic object detection,

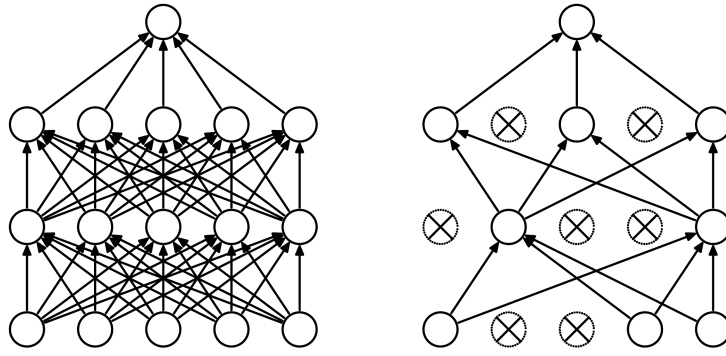


FIGURE 1.3: A neural network with standard dropout (Srivastava et al., 2014). (Left): A standard neural network with two hidden layers, one input and output layer. Note the connections between neurons of the two layers. Every neuron in a layer is connected to every other neuron in the layer below and above it. (Right): The effect of dropout is the dropping out of one or more neurons in the hidden layers, and as a result their connections are also dropped. The dropped neurons are shown as dotted circles with a cross in it.

is closely related to semantic image segmentation, which aims to assign each pixel in an image to a semantic class label. Unlike semantic segmentation, object instance segmentation, depicted in Figure 1.1 (d), aims to distinguish different instances of the same object class.

1.1.3 Dropout

Deep neural networks can suffer from overfitting due to their large number of parameters and the complexity of the learned features. This is especially so, when training with small datasets. As a result, the networks do not generalize well to work with unseen data. One obvious way to prevent overfitting, is to use a large training dataset. However, it is very time-consuming and expensive to create large, annotated datasets for supervised learning tasks.

Several approaches have been proposed to improve the performance of deep learning in the face of limited datasets. One approach is to average the weights of several networks to create a regularized model. However, this approach is computationally expensive as it requires many networks to be trained. Several other techniques like L_1 and L_2 regularization (Tian and Zhang, 2022), early stopping (Prechelt, 2002), augmentation (Shorten and Khoshgoftaar, 2019), batch normalization (Ioffe and Szegedy, 2015), and dropout (Srivastava et al., 2014), have also been proposed to solve overfitting. Of these, dropout methods have shown a lot of

promise in generalizing the models, using a simple and effective approach. As a result, they are now, well-integrated into all deep learning networks, as one or more dropout layers.

In dropout methods, the general concept while training, is to drop the neurons with a specified probability in the hidden layers of the network. The standard dropout, shown in Figure 1.3, was introduced by (Hinton et al., 2012). Dropout methods reduce co-adaptation and introduce sparsity, which help to generalize the model (Labach et al., 2019). As an example, if there are n neurons, then it is akin to training 2^n thin networks, since each neuron can be dropped with a uniform probability during the training phase. Hence, dropout can be seen as a stochastic regularization method. Dropout can also be conceptualised as adding noise to the network, while training (Vincent et al., 2008). This can be conceptualized as expanding the dataset by marginally adding noise. Adding noise to the dataset using a distribution obtained from the exponential family of function (Maaten et al., 2013) is another example of the noise-based approach. Gaussian dropout technique introduces Gaussian noise to the training set (Wang and Manning, 2013), (Srivastava et al., 2014). Bayesian models can also be approximated to dropout methods and the study (Gal and Ghahramani, 2016), provides theoretical evidence to link dropouts to Gaussian processes, and thus models uncertainty in deep learning.

1.1.3.1 Structural Changes

Some dropout methods involve structural changes, which include, dropping individual neurons, changing the connections between neurons, dropping a group of neurons, and applying Bayesian approaches or adaptive approaches (Li et al., 2022). Fast dropout (Wang and Manning, 2013), approximates the outputs from the dropout layers by sampling from a Gaussian distribution. Variational dropout (Kingma et al., 2015) and Monte Carlo dropout (Gal and Ghahramani, 2016) methods use the Bayesian approach of dropout. DropConnect (Wan et al., 2013), reduces the connectivity between the neurons rather than dropping the neurons. All these methods result in increasing the sparsity of the weights and biases in the network. Max-out method (Goodfellow et al., 2013) renders several outputs, out of which, the maximum value of the neurons is chosen. The adaptive approaches, such as jumpout (Wang et al., 2019c) and AS-dropout (Chen and Yi, 2021), obtain dropout parameters adaptively from a distribution, and an activation function, respectively. Another strategy uses auxiliary networks (Ba and Frey, 2013) to determine the dropout rate.

While the above methods deal with approaches to drop specific neurons for CNN networks, dropping groups of neurons has been shown to provide better regularization. For the CNNs, the regularization of the weights can be achieved by altering the network topology, or by removing the highly correlated blocks of features (Lee and Lee, 2020), (Ghiasi et al., 2018). Spatial dropout and DropBlock approaches, drop feature maps in CNN networks (Tompson et al., 2015), (Wu and Gu, 2015b). Spatial dropout also demonstrates that the information lost due to pooling can be recovered to some extent. Max-pooling dropout (Wu and Gu, 2015a), drops groups of neurons, before the information is passed to the pooling layer. In one study (Huang et al., 2016), layers of the network are removed during the training phase, thus modifying network topology. SwapOut (Singh et al., 2016), drops neurons by considering their residual connections, as well as their outputs and inputs.

1.1.3.2 Input Image Changes

Rather than making changes to the network to improve generalisation in the face of small datasets, an alternative is to expand the dataset through a technique known as augmentation. Typically, in these approaches, blocks of input images are erased, or noise added to them. Whilst augmentation can be 1-dimensional (1-D), 2-dimensional (2-D) or more, for computer vision applications, 2-D approaches are used. During the training phase, Cutout (DeVries and Taylor, 2017) and random erasing (Zhong et al., 2020) remove square and rectangular regions of the input image, respectively, while Hide and Seek method (Kumar Singh and Jae Lee, 2017) cuts out many regions in the same image. Mixup and Manifold Mixup (Verma et al., 2019), (Zhang et al., 2018) augment the image by interpolating the data at a feature level. Finally, combination methods, such as CutMix (Yun et al., 2019), combine Cutout methods with Mixup and (Walawalkar et al., 2020) combines Cutout with noise. In these methods, the pixels are removed and replaced with pixels from another image. The Figure 1.4 shows the effect of some of the augmentation strategies on the input image.

In dropout methods, there is a need to set the hyper-parameter for dropout, namely, the "dropout rate". The structural methods are complex and may lose some valuable information and cannot be easily adapted into existing networks. The augmentation methods increase the dataset but may not generalize well in all scenarios, as one requires domain knowledge of where and how to perform the augmentation.

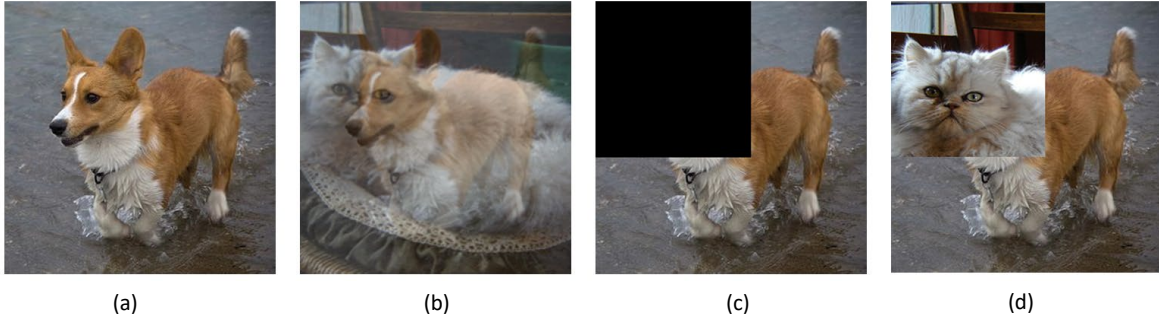


FIGURE 1.4: Modern augmentation strategies (Yun et al., 2019): (a) Original image; (b) Mixup augmentation; (c) Cutout or random-erase augmentation; (d) CutMix augmentation.

1.2 Data-Efficient Deep Learning

The most successful supervised machine learning algorithms discussed so far, including deep learning algorithms depend on large annotated datasets to achieve the required level of performance. There are many application scenarios where large amounts of fully annotated data is either not available or require an expensive and time-consuming process to create. This has motivated the research into data-efficient machine learning models which can achieve comparable results to fully supervised models with less training data and weaker supervision. Specifically, data-efficiency takes several forms: (1) weak supervision where the extent of data annotation is less comprehensive (2) unsupervised algorithms (3) generation of data for data-poor domains based on small examples (4) transfer of knowledge in the learned feature space rich-data domains into poor-data domains. In the following sections, we discuss the algorithms that have been proposed in the literature for data-efficient deep learning.

1.2.1 Weakly Supervised Methods

The lack of fully annotated large datasets to learn deep learning models, has led to the development of weakly supervised methods. These methods can be further classified as incomplete, inexact, and inaccurate supervision methods (Zhou, 2018). In the incomplete supervision method, only a small part of the dataset is labeled, while the rest remain unlabeled. In the inexact supervision, the data does not have detailed annotation. In inaccurate supervision, the data may have incorrect annotation (Zhou, 2018), (Nashaat et al., 2021).

In the inexact supervision method, the data annotations are coarsely specified. A well-known example of weakly supervised learning is "Multi-Instance Learning" (MIL) (Carbonneau et al., 2018). In the case of MIL, rather than annotating each image or instance of training data, a group of instances are labeled together. This group is called a bag of instances, and each bag is assigned a label. MIL can be conceptualized as a binary classification (or multi-class) problem at a bag level, rather than at an instance level, within the bag. Hence, regardless of the labels of individual instances within a bag, if there exists at least one instance with a positive label, then the bag is classified as positive. Another example of inexact supervision occurs in cell annotation, where each cell is annotated by a single dot, which can appear anywhere within the image of the cell¹.

In the inaccurate supervision method, the data annotations are sometimes wrong. This situation occurs in many common applications. For instance, crowd-sourcing may be used to annotate large datasets (Palotai et al., 2014). Due to either lack of domain knowledge, or human error, data may be inaccurately labeled. This type of error is called label noise, and several methods have been proposed to weed out this noise (Song et al., 2022). Most real-world datasets will contain some label noise. Hence, to address this problem, ensemble methods have been proposed to identify noisy labels. In one such method (Chen et al., 2018b), different classifiers vote on the noisy labels and using a weighted majority voting scheme, noisy labels are identified. Graph based methods (Muhlenbach et al., 2004) on the other hand, exploit the relationship between training data to remove noisy labels.

Since creating large, annotated datasets is an expensive exercise, and at times, impractical as well, methods using incomplete supervision have gained attention. A dataset used in incomplete supervision is only partially annotated. A sub-class of incomplete supervision which uses a human in the loop called active learning, aims to select the most appropriate samples from an unlabeled dataset, and pass it to a human annotator for labelling (Ren et al., 2021). The goal is to reduce the cost of labelling while maintaining optimal system performance. Deep learning algorithms have been quite successful in learning feature representations in various application scenarios but need significant amount of annotated data. Therefore, an obvious approach is to combine deep learning and active learning, as this will greatly expand their combined application potential by optimizing the process of querying the human for labelling unlabeled

¹Mitos and Atypia 14 <https://mitos-atypia-14.grand-challenge.org/Dataset/>

data. Uncertainty based strategies (Wang et al., 2016a), rely on an uncertainty ranking by the learner of the unlabeled data, and select the subset with the highest uncertainty. The Cost-Effective Active Learning (CEAL) (Wang et al., 2016a), during the training phase, feeds the human, unlabeled samples based on uncertainty of prediction. In the Bayesian Active Learning strategy (Tran et al., 2017), a separate Bayesian Convolutional Neural Network (Gal and Ghahramani, 2015) is used to determine the entropy of the learner’s prediction. The density-based approaches (Geifman and El-Yaniv, 2017) try to find an optimal subset to label which is representative of the entire unlabeled dataset.

Semi-supervised methods rely on some important assumptions such as generative models, clustering, manifold, co-training, and low-density separation assumptions, as specified in (Yang et al., 2022). The semi-supervised methods provide good predictions, based on the distribution of the unlabeled data. If the position of the unlabeled data is ambiguous within the labeled data, then the result from this method will be poor. Semi-supervised learning can be achieved using generative models, consistency regularization, graph methods, and pseudo-label methods (Yang et al., 2022). In the generative model (GAN) approach, training data is used to generate synthetic data with the same distribution as the input data. The newly generated data can be used as additional data, either for training, or classifier regularization (Yang et al., 2022). A number of GAN solutions have been proposed for semi-supervised classification, such as, CCGAN (Denton et al., 2016) and CatGAN (Chen et al., 2022). Consistency regularization is another method used for semi-supervised learning. In this approach, the consistency regularization, which is based on the manifold assumption, is applied to the loss function. An example of this is the teacher-student network, where the prediction consistency is maintained between the teacher and the student networks (Tarvainen and Valpola, 2017).

Although the semi-supervised methods discussed so far have achieved promising results, they only consider the unlabeled examples in isolation but ignore the associated relationships between these examples. Graph-based methods have been proposed in the literature to consider the underlying structure of the unlabeled dataset. If the unlabeled data were to be represented as a graph, then the nodes represent the individual samples of unlabeled data, and the edges represent the consistency of labels between nodes. Some methods (Luo et al., 2018) utilize the underlying structure of the data by building such a graph in the feature space, learned by the network. Then the network is expected to generate similar features for nodes from the same

class, while generating dissimilar features for nodes from different classes. The Reliable Edge Mining method (Chen et al., 2020) extends this idea further by constructing a graph, where each edge has two attributes: usefulness and certainty. During training, the certainty value is used to select the most reliable edges, which leads to faster training convergence. In another method (Isken et al., 2019), label propagation is performed to classify unlabeled data by constructing a nearest neighbour graph in the embedding feature space. Additionally, the method uses a certainty score per image and per class during training with unlabeled data.

Pseudo-label methods, shown in Figure 1.5, can be either disagreement methods or self-training methods. In the disagreement methods (Dong-DongChen and WeiGao, 2018), many networks are trained simultaneously, so that the unlabeled data can be labeled. The generation of pseudo-labels by self-training has shown promise in the disagreement methods. In the pseudo-label generation methods, pseudo-labels (Pseudo-Label, 2013), (Noroozi et al., 2018) are generated from unlabeled data and added to the network as training data. In both these cases, after obtaining the pseudo-label predictions on the unlabeled data, these are used for supervised training. The noisy-student approach uses an extension of pseudo-label methods with knowledge distillation in a teacher-student network (Xie et al., 2020).

Self-supervised approaches based on representation learning from a pre-task network (Kolesnikov et al., 2019), such as image rotation, develop a high-level semantic understanding, which is then used in downstream classifiers. A student-teacher network, where pseudo-labels generated by a teacher network are used to teach a student network has also been proposed (Pham et al., 2021). The authors call these meta pseudo-labels, since the teacher network is updated based on the feedback received from the student network. Semi-supervised training, with self-training, has been proposed to train an ensemble of robust classifiers (Wang et al., 2020). In this approach, a self-training network is trained using spatial and non-spatial transformations, to achieve effective feature representation.

Despite their advantages over fully supervised methods, weakly supervised methods still have challenges to overcome. Weak supervision, by definition, leads to less precise models compared to models trained with fine-grained supervision. Complex representations are difficult to learn and, as a result, the models may not generalize well on unseen data. Thus, domain knowledge is necessary to supplement weakly supervised learning, to overcome lack of precise or noisy labels.

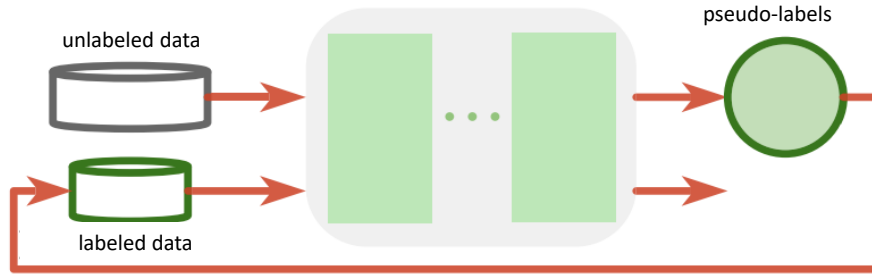


FIGURE 1.5: A generic pseudo-labeling method, where a network is trained with both labeled and unlabeled data and pseudo-labels are assigned to the unlabeled data, based on learned feature representations. The pseudo-labels are then used for future training phases for fully supervised training (Yang et al., 2022).

1.2.1.1 Unsupervised Learning

The field of unsupervised learning is a separate topic by itself, which includes diverse techniques such as, clustering (Xu and Wunsch, 2005), association of data (Cios et al., 2007), self-supervision (Ren et al., 2023), and dimensionality reduction (Sorzano et al., 2014). Furthermore, there has been lot of work in unsupervised DNNs. In the context of weakly supervised methods, there are instances, for example, in unsupervised domain adaptation, where there is full supervision on the labeled dataset and no supervision on the unlabeled dataset. Hence, in this example, there is a combination of both supervised and unsupervised learning. Hence, methods in which there is incomplete supervision, can be conceptualised as weakly supervised methods (Ren et al., 2023). The following section further discusses unsupervised domain adaptation.

1.2.2 Domain Adaptation

The supervised DNN models are robust when trained with large, annotated datasets. However, models that have been trained with the data collected in one domain, perform poorly when the target data is from a different domain with a different data distribution. For instance, if the model has been trained to detect vehicles using a dataset collected at daytime, such a model will not perform well under low-light conditions, such as late evening or night. One way to solve this problem is to gather more annotated training data from the target domain; in the example provided, night-time data. However, creating annotated datasets to cover target domains, encompassing all scenarios, can be both tedious and often impractical. Hence, domain

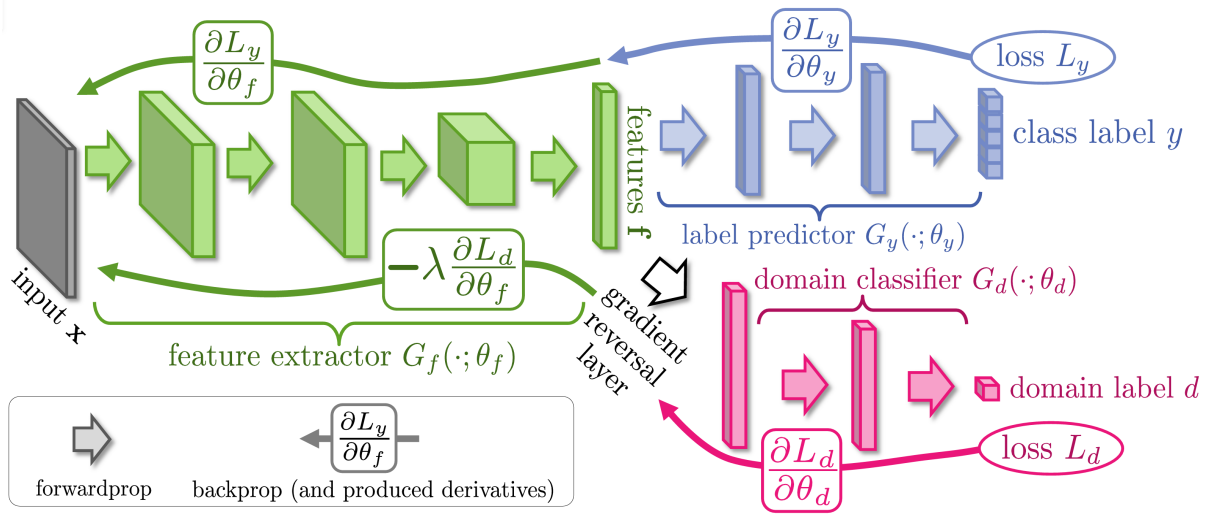


FIGURE 1.6: The proposed architecture includes a deep feature extractor (green) and a deep label predictor (blue), which together form a standard feed-forward architecture. Unsupervised domain adaptation is achieved by adding a domain classifier (red) connected to the feature extractor via a gradient reversal layer that multiplies the gradient by a certain negative constant during the backpropagation-based training (Ganin et al., 2016).

adaptation techniques are very useful to overcome this domain-shift problem. There are both weakly supervised, and unsupervised techniques for domain adaptation.

The domain adaptation methods proposed for unsupervised methods also work with weakly supervised methods, where only part of the target domain is annotated. Weakly supervised techniques use a labeled source dataset and a small percentage of a labeled target dataset. The rest of the target dataset is either unlabeled or may have inexact labels. For example, in one approach, a network is trained with weak, cross-domain supervision, where the source domain has instance-level annotations, and the target domain has image-level annotations (Inoue et al., 2018). In this method, the network learns to predict instance-level classification for the target domain using pseudo-labels.

Within the class of solutions which use domain adaptation, domain adversarial techniques (Ganin et al., 2016), (Saito et al., 2019), (Sindagi et al., 2020) have been applied successfully to bridge the domain gap. This technique has a label predictor to predict the source labels, and the domain discriminator to differentiate between source and target domains. The gradient reversal layer (GRL) (Ganin and Lempitsky, 2014) used in the network, is trained to obtain domain-invariant features. Figure 1.6 shows how a gradient reversal layer is incorporated into

an existing neural network. The domain adaptation using GRL, works on three components, the domain classifier, the label predictor, and the feature extractor. The goal of using gradient reversal is to align the learned-feature distributions of the source and target domains. As shown in the Figure 1.6, the features are extracted from the backbone network (Ganin et al., 2016). The feature extractor has two pathways. The first, is through the label predictor. The goal of the label predictor is to minimize the classification loss of the source dataset, in this case, minimizing the loss function $\frac{\partial L_y}{\partial \theta_y}$. The second pathway is through the domain classifier. The domain classifier is a binary classifier which distinguishes between source and target using the loss function $\frac{\partial L_d}{\partial \theta_d}$. Hence, the strategy, as given by equation Eq. (1.1), is to maximise the domain classifier loss $\frac{\partial L_d}{\partial \theta_d}$, while simultaneously minimizing the label classification loss $\frac{\partial L_y}{\partial \theta_f}$. This will guide the network to learn domain-invariant features by an adversarial training approach. This is obtained by arriving at a saddle point for the parameters in the below equation Eq. (1.1) (Ganin et al., 2016).

$$\sum_{i=1, d_i=0}^N L_y^i(\theta_f, \theta_y) - \lambda \sum_{i=1}^N L_d^i(\theta_f, \theta_d). \quad (1.1)$$

In the context of stochastic gradient descent, this saddle point can be achieved by introducing a gradient reversal layer between the feature extractor and the domain classifier (Ganin et al., 2016). The gradient reversal layer has no parameters except λ . In the forward pass, the gradient reversal layer performs an identity function, and during back-propagation, the gradient is multiplied by $-\lambda$, which is a constant multiplier. Thus, the gradient reversal layer helps the network to learn domain adaptive features, though it may not be ideal when the domain gap is large. It also has the risk of overfitting to domain-specific features.

The self-supervised approaches, which are a type of unsupervised learning, (Liu et al., 2021a), (Jaiswal et al., 2020) seek to learn from the unlabeled data, automatically, without any supervision. Though pseudo-label, self-training approaches have been used in semi-supervised learning, recently, they have also been used in unsupervised approaches as well (Zou et al., 2019). These approaches generate pseudo-labels from the unlabeled dataset and apply the labels for supervised training. Several approaches (Pham et al., 2021), (Inoue et al., 2018), (RoyChowdhury et al., 2019) have been proposed for self-supervised training using pseudo-labels. A common strategy followed by these methods is to use a source domain dataset, which is labeled, and as

the training progresses, generate pseudo-labels for the target domain. The pseudo-labels are then used as input in training along with the labeled data for performing supervised training. These methods have the drawback of introducing noisy labels in the training. There are several methods introduced to mitigate this noise (Liu et al., 2021b), (Zou et al., 2019).

Furthermore, there are domain adaptation methods described in the literature (Rodriguez and Mikolajczyk, 2019), (Zhang et al., 2019) in which, target images are translated so that they are similar to the source images. The generative adversarial methods (GANs) are adapted in the literature to obtain this translation. Some of these methods are also used in conjunction with other methods, such as those obtaining domain invariant features (Zhang et al., 2019). The GAN methods provide a translation from the source domain to the target domain. However, sometimes, these translations are not robust. Hence, there are methods which transform source domains to several intermediate domains, to achieve domain invariance. For instance, in one such method (Kim et al., 2019), Faster R-CNN is trained using three datasets, namely, source domain, target domain, and the transformed intermediate domains of the source domain. The intermediate-styled domains are derived with the right constraints from the source domain, in such a way that the model can be generalized. In a similar manner, the study (Rodriguez and Mikolajczyk, 2019) uses an intermediate-style approach, by training a styled network. Intermediate domains are complex, and if not aligned accurately, may introduce additional noise. Furthermore, GANs are difficult to train and suffer from mode collapse.

Graph-based approaches and multi-task learning approaches have also been proposed for domain adaptation. Graph-based approaches try to model inter- and intra-object relationships during training. The Mean Teacher with Object Relations (MTOR) (Cai et al., 2019) method, explores a novel student teacher network, using a region proposal network as the base, where the teacher and the student are trained with different augmentations of the same image. During training for object detection, a graph is constructed using the region proposals from both the teacher and student, based on the feature similarity of pairs of regions between the two networks. Along with the supervised detection loss, the method proposes three consistency regularizations, at region-level to align region-level predictions, at graph-level to match graph structure between teacher and student, and within a graph to enhance the similarity between regions of the same class. The consistency regularization helps the network learn invariant features across domains. The Deep Reconstruction-Classification Network (DRCN) (Ghifary

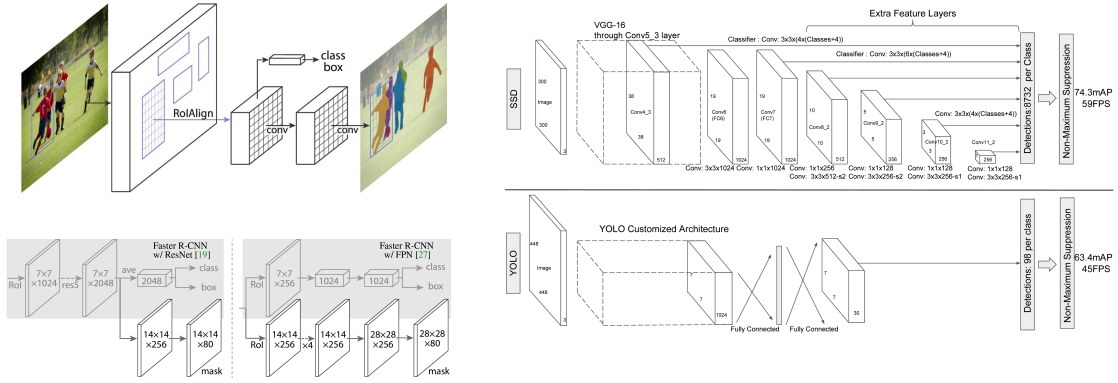


FIGURE 1.7: (left) The state-of-the-art dual stage Mask R-CNN network which is based on Faster R-CNN (He et al., 2017), (right) single-stage monolithic networks YOLO and SSD (Liu et al., 2016b).

et al., 2016), uses a technique similar to multi-task learning for learning domain invariant features. The goal of multi-task learning is that, learning auxiliary tasks can help the main task be learned better. In this spirit, the authors of this study hypothesize that, domain adaptation can be achieved by learning the label distribution in the source domain and the data distribution in the target domain, and hence, the authors propose a convolutional network that jointly learns supervised source label prediction and unsupervised target data reconstruction.

1.3 Applications

1.3.1 Computer Vision Applications

Over the years, in the field of computer vision, several important application domains have emerged, such as generic object detection, object counting, face recognition, action and activity recognition, and human pose estimation. However, the application areas relevant to this thesis, namely, person re-identification, object detection and object counting, are described briefly.

1.3.1.1 Object Detection Frameworks

Detecting a wide variety of objects, and determining their exact location within a complex scene, has many applications. As different detection tasks have different objectives and constraints, their difficulties may vary. Some common challenges in computer vision tasks include

objects under different viewpoints, illuminations, and intraclass variations. Additional challenges in object detection include, but are not limited to, object rotation and scale changes (e.g. small objects), accurate object localization, dense and occluded object detection, and speed-up of detection. These varying objectives have led to the development of different object detection frameworks with different applications and use cases.

Object detection frameworks can be classified into dual-stage, region-based frameworks, and unified, single-stage pipelines. In two-stage methods, category-independent region proposals are first generated from an image. Region-based CNN (R-CNN) (Girshick et al., 2014), proposes a method of generic object detection using CNNs. R-CNN integrates AlexNet (Krizhevsky et al., 2012) with a region proposal, based on selective search (Uijlings et al., 2013). R-CNN has several drawbacks. Chief among them, is that each stage of the multi-stage training pipeline needs to be individually trained and optimized. The drawback of RC-NN has led to further innovations, leading to a number of improved detection frameworks such as SPPNet (He et al., 2015), Fast R-CNN (Girshick, 2015), Faster R-CNN (Ren et al., 2015), and Mask R-CNN (He et al., 2017), which train the two-stage pipeline in an integrated fashion. Mask R-CNN combines generic object detection with semantic image segmentation which results in prediction of pixel-level masks for individual object instances found in the image.

In contrast, unified pipelines refer to architectures that directly predict class probabilities and bounding box offsets from full images. They employ a single, feed-forward CNN in a monolithic setting, which does not involve region proposal generation, post classification, or feature resampling, encapsulating all computation in a single network. Since the whole pipeline is a single network, it can be optimized end-to-end, directly, on detection performance. "You Only Look Once" (YOLO) (Redmon et al., 2016) is a unified detector which predicts bounding boxes and associated class probabilities, directly from the image. Single Shot Detector (SSD) (Liu et al., 2016b), achieves real-time speed without sacrificing detection accuracy, and is faster than YOLO with an accuracy comparable with region-based detectors, such as Faster R-CNN. SSD effectively combines ideas from region proposal networks (RPN) Faster R-CNN, YOLO, and multi-scale convolutional features (Lin et al., 2017), to achieve fast detection speed, while still retaining high detection quality. Like YOLO, SSD predicts a fixed number of bounding boxes and scores, followed by a non-max suppression (NMS) step, to produce the final detection. Figure 1.7 depicts Mask R-CNN, YOLO, and SSD frameworks.

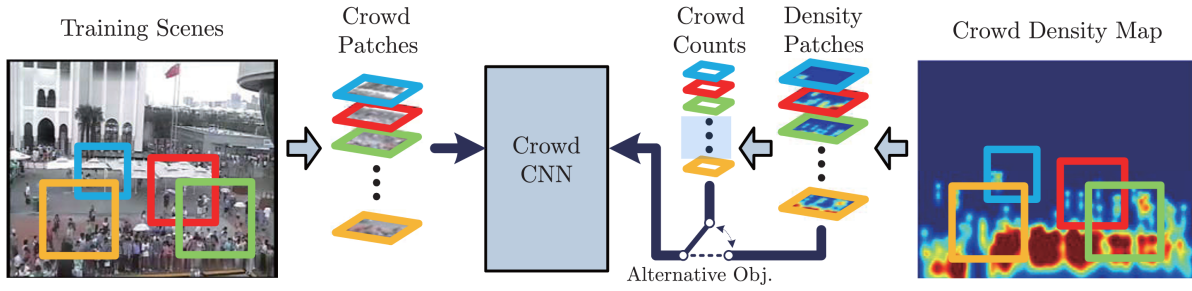


FIGURE 1.8: Density-map based Crowd Scene Counting Method (Zhang et al., 2015). In this method patches on images containing people, are used to train a CNN with two training objectives: (1) counting the crowd in the patch; (2) generating a density map for the patch. The multi-task training proposed here improves the overall accuracy of the crowd counting.

1.3.1.2 Object Counting

Object counting in machine learning provides an automatic way of estimating the number of objects in images or video. This is useful in many applications, such as microscopy cell counting, people or crowd counting for surveillance, and counting wild-life or trees with aerial imaging (Ranjan et al., 2021). Counting has many challenges, such as high-density objects, occlusions, size variability, overlapping objects, difficult lighting conditions, and distracting backgrounds.

The approaches in counting can be either class-specific or class-agnostic. Furthermore, these approaches can be regression-based or detection-based. In class-specific counting, knowledge of the object to be counted is required. In these methods, extracted feature representations are used for counting (Chan et al., 2008), (Chen et al., 2012) or by using appearance of objects by considering large patches around them (Mundhenk et al., 2016). Regression-based approaches use a density map for counting objects. Figure 1.8 shows the use of a density map for crowd counting (Zhang et al., 2015), (Lempitsky and Zisserman, 2010). In these approaches to crowd counting, an integral of the pixel weights of the image is determined to count the objects in the image patch. Similarly, (Zhang et al., 2015) proposes a multi-task solution, where the network learns both the crowd density as well as crowd count. Furthermore, there are regression-based approaches, where fewer object annotations (weakly supervised) are used for counting (Wang et al., 2019a), (Chen et al., 2021). Focus on positive labels and label propagation is used by (Wang et al., 2019a) for handling incomplete labels. Self-attention methods, using transformer have recently been applied for crowd-counting (Liang et al., 2022) while, (Meng et al., 2021) proposes a mean teacher-student network to predict a density map and estimate its uncertainty.

In a more recent study on crowd counting, a density-based approach using pseudo-labels with multi-task learning has been proposed (Zhu et al., 2023).

The detection-based approaches by (Amato et al., 2019), (Laradji et al., 2018), (Ciampi et al., 2018) use detection or segmentation (He et al., 2017) to detect and count the objects. Instance segmentation or bounding box-detection methods are used for detecting the objects. The study (Hsieh et al., 2017), uses a region proposal approach with spatial weights to locate and count the objects. Counting-DETR (Nguyen et al., 2022) and region-proposal counting method (Ranjan and Nguyen, 2022), use an object or instance detection method to predict bounding boxes with uncertainty estimation. This uncertainty calculation is required, as pseudo-bounding boxes are generated from the first detection stage.

There are also class-agnostic methods of object counting. In the class-agnostic methods, generally an exemplar is selected in the image, for which the count is obtained. In object counting, the representative object that is to be counted in the image, is called an exemplar. These can be conceptualised as few shot methods, as the methods also work with unknown classes. These methods may use detection or regression approaches for object counting. Sometimes, exemplars are defined outside the image, separately (Xu et al., 2023). The objects in the image are matched against the exemplars and counted. GMN (Lu et al., 2019), uses this approach of density or detection-based counting with a domain-adaptive network. FamNet (Ranjan et al., 2021) matches the exemplar features with the feature correlation matrix of the image. LaoNet (Lin et al., 2021) learns both intra- and inter-relations and incorporates multi-scale feature aggregation. A zero-shot counting approach (Xu et al., 2023) provides an exemplar-free method to count objects, by using density maps to ascertain the quality of the exemplars.

The disadvantages of regression-based methods are that they lack the ability to identify the location of individual objects. Weakly supervised methods perform poorly, if the features of the exemplars are similar to other similar objects in the image. Furthermore, if the objects are overlapping, then it is difficult for the method to distinguish the objects. Also, they perform poorly if the objects are of different shapes and sizes. The detection-based methods using instance segmentation, rely on per-pixel annotation, which is time-consuming and expensive. But if the annotation is coarse (e.g. bounding box instead of pixel-level), then it suffers from the same disadvantage as the regression-based techniques. Hence, counting techniques still have

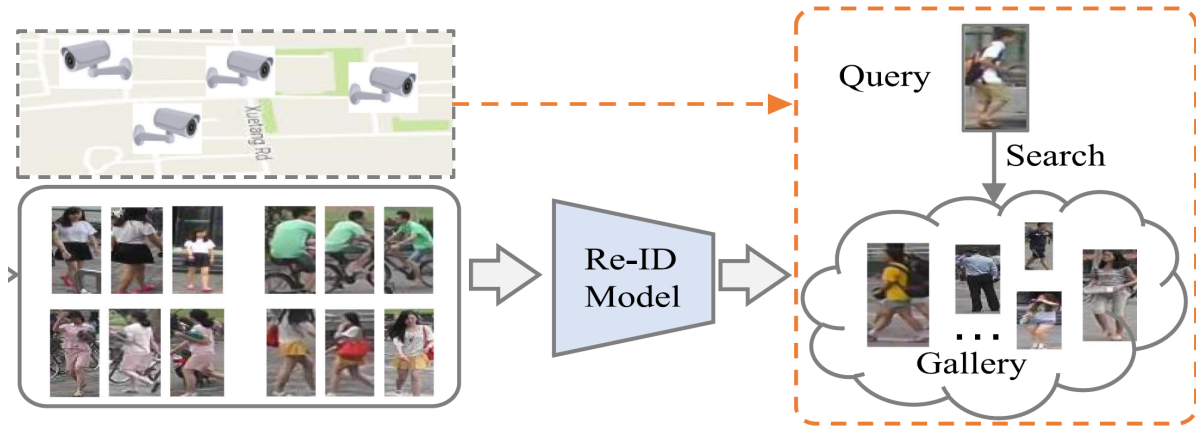


FIGURE 1.9: Person ReID representation learning from images captured from multiple cameras and matching a query image of a specific person with a gallery (Ye et al., 2021).

major challenges to overcome when there is a high density of objects, presence of overlapping objects, and changes in conditions such as lighting or other environmental factors.

1.3.1.3 Person Re-Identification

In the last two decades, there has been an acute need for public safety, which has led to an increase in the number of surveillance cameras. This has led to a unique class of applications involving detection of persons across non-overlapping camera views. This is called person re-identification, (ReID) and is a challenging problem in itself. Solving the person ReID problem is imperative in intelligent surveillance systems, and has led to significant research activity, due to its practical importance. Hence, person ReID has been studied quite extensively, and this has led to many advances in the field of deep learning.

Person ReID can be conceptualised as finding whether a person of interest in one camera view, has appeared at a different or at the same time, in another camera view. It can be cast as a specific, person-retrieval problem, across non-overlapping cameras (Chen et al., 2017), (Zheng et al., 2016b). Given a person-of-interest, namely, the query, the goal of ReID is to determine whether this person has appeared in another place at a distinct time captured by a different camera, or even the same camera at a different time (Gheissari et al., 2006). This set of images is termed as a "gallery". Figure 1.9 illustrates the overall Person ReID process. The query can be represented by an image (Almazan et al., 2018), (Zheng et al., 2015), (Martinel et al., 2019),

a video sequence (Wang et al., 2014b), (Zheng et al., 2016a), or even a text description (Li et al., 2017b).

Before the application of deep learning for person ReID, Metric learning has been extensively studied by learning either a Mahalanobis distance function (Koestinger et al., 2012), (Zheng et al., 2011) or a projection matrix (Liao and Li, 2015). With the advances of deep learning, person ReID approaches in a fully-supervised setting, now use global feature representation learning. The supervised ReID approaches use DNNs, which were initially designed for classification tasks (Simonyan and Zisserman, 2014), (He et al., 2016) to learn these feature representations. All supervised ReID learning approaches, use one of identity loss (Zheng et al., 2017a), verification loss (Varior et al., 2016), (Deng et al., 2018), or triplet loss (Yi et al., 2014), to train deep neural networks to learn discriminative feature representations. Identity loss treats the training process of person ReID as an image classification problem, where each identity is a distinct class. Triplet Loss treats the ReID model training process as a retrieval ranking problem. The basic idea behind triplet loss is that the distance between the positive pair (images of same identity) should be smaller than the negative pair (image of different identities) by a pre-defined margin. Verification loss optimizes the pairwise relationship, either with a contrastive loss (Varior et al., 2016), (Deng et al., 2018) or a binary verification loss (Li et al., 2014), (Zheng et al., 2017b). The verification is often combined with the identity loss to improve the performance (Li et al., 2014), (Zheng et al., 2017b), (Chen et al., 2018a).

Attention schemes have also been widely studied in literature to enhance representation learning (Yang et al., 2019). Attention can be within a person's image using pixel-level attention (Li et al., 2018b), spatial information (Chen et al., 2019a), and multiple person images with intra- and inter-sequence attention (Si et al., 2018). Unsupervised domain adaptation (UDA) transfers the knowledge on a labeled source dataset to the unlabeled target dataset (Ma et al., 2013). Due to the large domain shift and full supervision in the source dataset, it is another popular approach for unsupervised ReID where target dataset is unlabeled. The success of UDA has led to unsupervised ReID receiving increasing attention in recent years. Using GANs to transfer the source domain images to target-domain style is a popular approach within UDA ReID [(Wei et al., 2018), (Chen et al., 2019b), (Wei et al., 2018), (Li et al., 2019)]. There are also approaches in the literature based on use of pseudo-labels and ensemble of networks [(Ge et al., 2020a), (Fu et al., 2019), (Wang et al., 2022), (French et al., 2017), (Ge et al., 2020b), (Dai et al.,

2021)]. Despite the advances in person ReID, there still remain many challenges, including re-identification in cross-domain setting, handling occlusions, different camera configurations and pose changes.

1.3.2 Medical Image Applications

Cancer continues to be a major cause of death in the world (Torre et al., 2015). Thus, fighting against it is a challenge faced by the research community as well as clinicians (Siegel et al., 2021). Among cancers, Breast cancer is among the top three, according to World Health Organization². Medical imaging is a very important technique for early cancer detection and diagnosis. However, manual interpretation of images is tedious, time-consuming and prone to human bias and errors. Hence, in the last few decades, computer-based systems have been assisting doctors in interpreting medical images (Doi, 2007). Analysis of medical images require tackling some fundamental problems such as detection and classification (e.g. mitosis cells, tumors) (Ludovic et al., 2013) and well counting cells (e.g. blood cells in leukaemia) (Suryani et al., 2015). With the advances in machine learning and deep learning, medical image analysis has shifted towards use of CNNs for different medical application scenarios. In the context of this thesis, this section describes medical applications, relevant to mitosis cell detection and classification in histopathology images, as well as cell counting in microscopy images.

1.3.2.1 Mitosis Cell Recognition

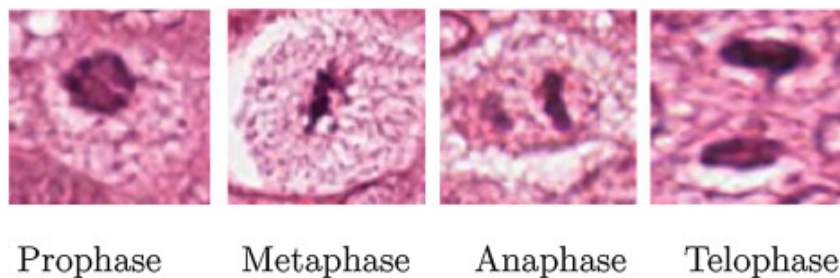


FIGURE 1.10: Different phases of Mitosis Cells (left to right): Prophase, Metaphase, Anaphase, and Telophase (Beevi et al., 2019). Note that the shape of a nucleus varies as the cell goes through the 4 stages, and in the last phase where there are two divided nuclei, the count of cells is still one. These morphological variations in the cell nucleus poses a great challenge for automatic cell counting.

²<https://www.who.int/news-room/fact-sheets/detail/breast-cancer>

In digital pathology, whole slide scanners are used for analyzing tissue images. Analysis (Doi, 2007) of cells in whole-slide images is used for cancer grading and prognosis. This is done by recognizing and counting the mitotic cells. The tissue biopsy to be analyzed, is stained, fixed and also sectioned differently. Hence, there is a lot of variability demonstrated in texture, structure, and background in the cells. Cell detection is also very tedious and time-consuming. Mitotic cells closely resemble other cells (mimics), as depicted in Figure 1.11, which makes it very difficult to identify them. Mitotic cells also have four different phases, as depicted in Figure 1.10. Each phase has a different cell morphology, and hence makes it challenging to identify (Wilkins and Holliday, 2009). Additionally, these cells are sparsely present.

Traditional methods, using hand-crafted features, which analyze the textural, morphological, and structural aspects of the cells (H. Irshad, 2013), (Thiran and Macq, 1996), (Choi et al., 1997), do not perform very well due to variability in tissue preparation. These methods do not accurately distinguish the mitotic cells from their mimics. Hence, deep learning solutions provide an advantage over these traditional methods.

Stain normalization converts an image into another, so that the visual characteristics of the image are better for analysis (Salvi et al., 2021). Some methods focus on separating the colour from pixel intensity, and perform stain normalization by transferring mean colour information or colour histogram information, from source to target images (Macenko et al., 2009), (Ruifrok et al., 2001). In deep learning, stain normalization has been done using GANs, transfer of RGB colour from source to target, and adaptive colour normalization and distribution (Cho et al., 2017), (Shaban et al., 2019), (Bug et al., 2017), (Zheng et al., 2019). In many DNNs, stain normalization has been used (Veta et al., 2014), as it provides better accuracy (Salvi et al., 2021).

Deep learning mitosis cell detection and classification can be broadly classified into pixel-based, region-based, and cascade approaches. Pixel-based approaches, consider patches across the whole image with a fixed bounding box around the mitosis cell (Cireşan et al., 2013). This method is computationally expensive. Region-based approaches can be single-phase (Sebai et al., 2020) or two-phase. Two-phase approaches use region-proposal networks (Li et al., 2018a), where the first step is used for detection, and the next, to eliminate false positives (Li et al., 2018c), (Chen et al., 2016), (Li et al., 2018a). Extending the two-phase region-proposal network, a hybrid approach has been proposed using a combination of learned deep features and

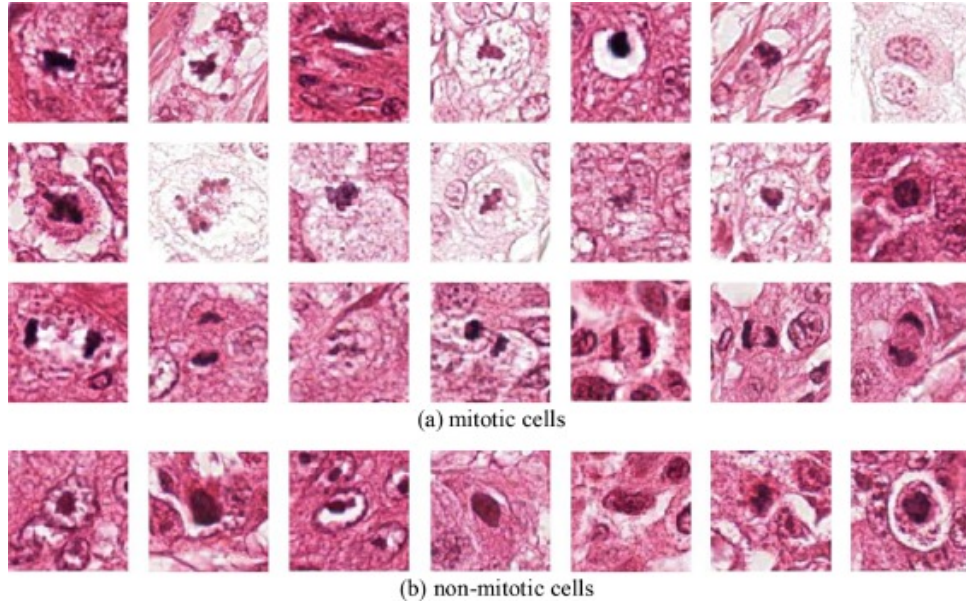


FIGURE 1.11: Mitosis Cells and its mimics (Li et al., 2018a): (a) The first three rows, are mitosis cells and the third row are cells undergoing telophase and have two nuclei; (b) Mimics of mitosis cells are in the fourth row. They are similar to mitosis cells in appearance.

hand-crafted features (Saha et al., 2018), (Wang et al., 2014a). The region-based approaches require good quality bounding-box annotation. Furthermore, the cascade methods rely on two-step processes, which are time-consuming. Segmentation networks (Li et al., 2018a) have also been used to generate masks for weak labels. To cater for different stains, as shown in Figure 1.12, cross-domain deep learning methods have been used, as described in the above section.

Even though there are a lot of advances in mitosis detection and classification for cancer diagnosis, many challenges remain, including, different staining methods, distinguishing mitotic cells from mimics, and morphological differences in mitotic cells.

1.3.2.2 Microscopy Cell Counting

Microscopy cell counting has been useful for medical condition diagnosis, in understanding the impact of drugs on cell number, type and structure changes, and in biological studies and experimentation (Mukherjee et al., 2017). Cell counting is very challenging, due to their large numbers as seen in DCC (Marsden et al., 2018), and MBM (Paul Cohen et al., 2017), or sometimes being very sparse like mitotic cells (Roux et al., 2013), differentiability with their mimics,

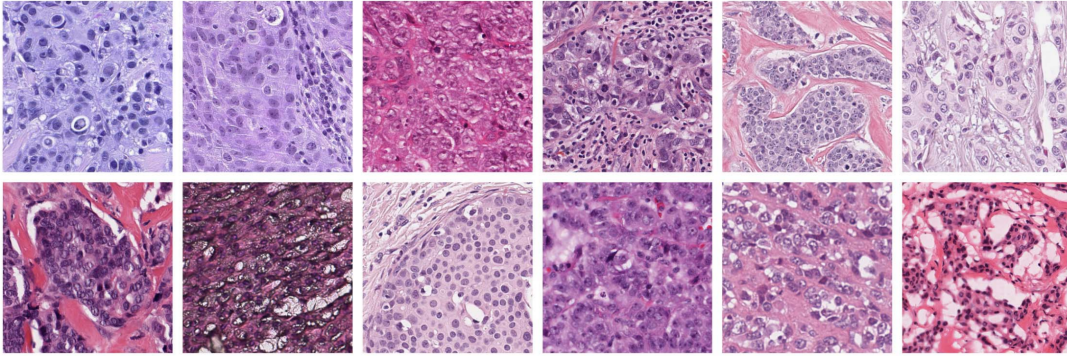


FIGURE 1.12: Breast tissue samples stained with hematoxylin and eosin from the Tumor Proliferation Assessment Challenge (TUPAC) dataset (Tellez et al., 2018). The images come from different laboratories and as a result have different staining.

changes in morphology, and different staining conditions and tissue preparation. Additionally, cells may be overlapping and have diffuse boundaries.

In the literature, cell-counting methods are categorized into two main groups, regression methods (Paul Cohen et al., 2017), (Jiang and Yu, 2020), (Xie et al., 2018), (Rad et al., 2019), and detection methods (Chen and Chefd’Hotel, 2014), (Liu and Yang, 2017), (Yusro et al., 2023), (Rad et al., 2018). In detection methods, instance segmentation and bounding box prediction methods are used. The segmentation methods require pixel-level annotations, which is a very time-consuming and tedious process, though it has the advantage of better localised regions. The bounding-box detection methods require weak annotations, such as dot annotations. These detection methods are coarser but have the advantage of a less expensive annotation process. However, this process suffers from decreased counting accuracy. These detection methods have a major disadvantage, as the detection performance is affected by occlusions and cell morphological changes.

Regression-based cells approaches do not rely on detections in the image. They provide an estimate of the cell count by using density maps, which are obtained by the sum of the pixel values. The regression-based methods require only dot annotation. The regression methods of cell counting can be further classified into supervised and weakly supervised methods. The weakly supervised method reduces the annotation overhead but relies on domain knowledge for label propagation from the labeled to the unlabeled dataset.

Cell counting methods rely on exemplars, which are objects, representative of a specific class, which is used for counting. Exemplars can be defined externally, or within the image. This can be manually selected by a human, or automatically determined. These exemplars should encapsulate cell morphological and texture variations and should have an uncluttered background. The quality of the exemplars play an important role in cell counting accuracy. Microscopy cell counting still has many unresolved challenges, such as impracticality of creating fully annotated datasets and lack of objective methods for defining good quality exemplars, when using weakly annotated datasets.

1.4 Evaluation Metrics

The effectiveness of the advanced learning-based methods for computer vision applications is measured using metrics which depend on the underlying learning task. As relevant to this thesis, the fundamental learning tasks which we have studied are: (1) binary classification, (2) object detection, (3) object counting, and (4) person re-identification. In this section, we describe the metrics used to evaluate the machine learning tasks as described in the literature, which are relevant to this thesis.

1.4.1 Metrics for Binary Classification

TABLE 1.1: Confusion Matrix.

		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

For binary classification problems, the evaluation uses a confusion matrix as shown in Table 1.1 (Naidu et al., 2023). True Positive (TP) represents correct positive classifications, False Positive (FP) represents incorrect positive classifications, False Negative (FN) represents missed positive classifications, and True Negative (TN) represents correct negative classifications. The row of the table represents the predicted class, while the column represents the actual class.

Based on the definition of the confusion matrix, the following metrics are defined:

$$Precision(P) = \frac{TP}{TP + FP}, \quad (1.2)$$

$$Recall(R) = \frac{TP}{TP + FN}, \quad (1.3)$$

$$Accuracy(acc) = \frac{TP + TN}{TP + FP + TN + FN}, \quad (1.4)$$

$$F_1 = \frac{2 \times P \times R}{P + R}. \quad (1.5)$$

The $Precision(P)$ is measure of correct positive predictions by a classification model as defined by the equation Eq. (1.2). A higher precision value indicates fewer false positives and better accuracy in identifying true positive instances among the predicted positives. The $Recall(R)$ measures a classification model's ability to identify all positive instances correctly as given by equation Eq. (1.3). A higher recall value indicates that false negatives are minimized and true positives are maximised. $Accuracy(acc)$ is a classification model's overall performance in terms of the correctness of both positive and negative predictions as given by equation Eq. (1.4). Accuracy is a reasonably good metric to use for datasets which are balanced, i.e. there are approximately an equal number of positive and negative samples in the dataset. Finally, the F_1 Score as shown in equation Eq. (1.5) considers the harmonic mean of precision and recall and is a good indicator of classification accuracy (Powers, 2020).

1.4.2 Metrics for Object Detection

The key metrics used for evaluating the accuracy of object detection models are Average Precision (AP) and Mean Average Precision (mAP) (Everingham et al., 2010). For object detection models, the outputs are a bounding box, a classification category for the bounding box, a confidence score, and optionally an instance segmentation mask. In this context, a true positive (TP) is a correct detection of a bounding box, a false positive (FP) is a bounding box which is not enclosing any object, and a false negative (FN) is an undetected object. Note that the

concept of true negative does not exist in the case of object detection. To determine whether a detection is a correct one or not, an Intersection Over Union (IOU) is used, which is based on the Jaccard Index (Zhong et al., 2017), a measure of similarity between two sets of data. The IOU is defined as follows:

$$IOU = \frac{area(B_p \cap B_g)}{area(B_p \cup B_g)}, \quad (1.6)$$

where B_g is the ground truth bound box and the B_p is the predicted bounding box. Furthermore, the IOU is compared with a pre-determined threshold t and a bounding box is classified as a correct if $IOU \geq t$ and if $IOU < t$ the detection is considered as incorrect. Following on from the definition of IOU , a detection is determined to be one of a true positive (TP), a false positive (FP) or a false negative (FN). This will allow the calculation of both precision (P) and recall (R) as given by the equations Eq. (1.2) and Eq. (1.3) respectively. In object detection tasks, for each detection, a confidence score is also predicted by the model. Precision and recall values at different confidence thresholds can be used to plot a precision-recall curve. The area under the precision-recall curve is a measure of model performance and is known as Average Precision (AP). Average Precision is applicable to object detection tasks where there is only one object category to predict. A further metric known as Mean Average Precision (mAP) is used for multi-class object detection tasks. The mAP is the mean of APs determined for each object category. The Average Precision (AP) is defined as follows:

$$AP = \frac{1}{11} \sum_{R \in \{0, 0.1, \dots, 0.9, 1\}} P_i(R), \quad (1.7)$$

where

$$P_i(R) = \max_{\tilde{R}: \tilde{R} \geq R} P(\tilde{R}). \quad (1.8)$$

In the above definition of AP as defined by equations Eq. (1.7) and Eq. (1.8), the maximum precision $P_i(R)$ whose recall value is greater than or equal to R is considered. Furthermore, the maximum precision is determined at 11 different equally spaced recall values.

The mAP follows from the definition of AP by considering AP over all object categories and is defined as follows:

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_c, \quad (1.9)$$

where AP_c is the average precision of the c^{th} class, and N is the total number of classes.

As part of object detection, if instance segmentation is also performed, then pixel-level masks are predicted for each detected objects. In this case, another metric called the DICE coefficient (Dice, 1945) becomes relevant for evaluating such models. In the field of medical image analysis the DICE metric is widely used. It is defined as follows:

$$DICE = \frac{2 \times |M_p \cap M_g|}{|M_p| + |M_g|}, \quad (1.10)$$

where M_p is the pixel-level mask that is predicted and M_g is the pixel-level ground truth mask. Furthermore, $|M|$ represents the cardinality, which in this case corresponds to number of pixels.

1.4.3 Metrics for Person Re-Identification

In the context of person re-identification (ReID), the Cumulative Matching Characteristic (Zheng et al., 2016b) (CMC) is one of the evaluation metrics used to assess the effectiveness the ReID model. The CMC metric evaluates the system's ability to correctly match a query image with a set of gallery images, by determining the ranks of the matching gallery images. CMC can also be conceptualized as the probability of a correct match within top- k gallery images. The top- k matches are determined using a distance measure such as Euclidean distance or a cosine distance.

Associated with CMC is the CMC curve, which plots the cumulative matching rate against the rank, where the x-axis is the rank (starting from rank-1, 2, 3 etc.), and the y-axis is the probability of finding the correct match within that rank. Thus, the curve can be used to report RANK-1, RANK-5 and RANK-10 (or other similar ranks) probabilities.

The CMC metric when retrieving top- k gallery images for a given query image (CMC_k), is defined as follows:

$$CMC_k = \frac{1}{Q} \sum_{i=1}^Q f(b_i, k), \quad (1.11)$$

where b_i is i^{th} query image, Q is the total number of query images, and $f(b_i; k)$ is an indicator function which equals to 1 when there is at least one matching gallery image within the top- k rank, and 0 otherwise.

From the equation Eq. (1.13), RANK-1, RANK-5, and RANK-10 metrics can be determined. In addition to CMC, the Mean Average Precision (mAP) (Zheng et al., 2015) has also been used to evaluate person ReID models. Firstly, the AP (AP_q) for each query image is determined, following which mAP (mAP_{reid}) is obtained as follows:

$$AP_q = \frac{1}{N_g} \sum_{k=1}^N P(k) \times R(k), \quad (1.12)$$

$$mAP_{reid} = \frac{1}{Q} \sum_{i=1}^Q AP_i, \quad (1.13)$$

where for a given query image q , $P(k)$ is the precision at a particular rank k , $R(k)$ is the recall indicator which is 1 if the k^{th} gallery image matches the query image q , N_g is the number of gallery images that match the query image as per the ground truth, and Q is the total number of query images.

1.4.4 Metrics for Object Counting

For object counting (especially cell counting) in medical image analysis, the Mean Absolute Error (MAE), Mean-Squared Error (MSE) and Root Mean-Squared Error (RMSE) are the three commonly used metrics (Ranjan et al., 2021), (Chai and Draxler, 2014). When these metrics are used, the object counting is typically cast as a regression problem.

The metrics are defined as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|, \quad (1.14)$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad (1.15)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \quad (1.16)$$

where y_i is the number of actual objects (e.g. cells) in the i^{th} image, $\hat{y}_i$ is the number of predicted objects in the i^{th} image, and n is the number of objects. Note that both MAE (equation Eq. (1.14)) and RMSE (equation Eq. (1.16)) are in the same units as the object count, whereas MSE (equation Eq. (1.15)) is a squared-error.

1.5 Challenges

The existing supervised and weakly supervised methods have shown lot of promise and have achieved state-of-the-art performance in many applications. However, there are many challenges that remain unsolved, including imprecise localisation, lack of inadequate source and target alignment, lack of adaptive generalization ability, and datasets having weak or incomplete annotation. In the following four sections, we briefly summarise these challenges in learning deep learning models for visual content analysis tasks.

1.5.1 Imprecise Localisation

In the Sections 1.1.2 and 1.1.1, we described methods for detecting and localising objects using bounding boxes, as well as instance-level masks. The instance-based segmentation methods, such as Mask R-CNN (He et al., 2017), provide a good localisation and have outperformed bounding-box based methods, such as Faster R-CNN (Ren et al., 2015) and YOLO (Redmon et al., 2016). This is because detection approaches without masks, do not provide accurate localisation, as they do not capture the morphology of the object. Additionally, a bounding-box

might include parts of other objects and background, forming an inexact annotation. As a result, existing methods that rely on weaker bounding box annotations, lead to inaccuracies in localization. These inaccuracies have a cascading effect on further stages of the training pipeline, such as object classification. For example, an improperly localised region can make the learned feature representations noisy. This, in-turn, reduces the accuracy of detection and classification. Instance segmentation methods also have further challenges due to lighting changes, cluttered background, and object morphological variations between training and testing datasets. Finally, preparing datasets with pixel-level annotations, is complex in scenarios where there are small and overlapping objects. Such scenarios are common in medical applications dealing with cell detection.

1.5.2 Lack of Adaptive Generalization

Neural networks typically have a large number of parameters. One of the problems with having a large number of parameters, is overfitting. This means that the trained models do not generalize well, and hence, will perform poorly on unseen data. Dropout is a popular and well understood regularization technique used for reducing co-dependence of the neurons leading to a more generalized deep learning model. Standard dropout requires hyper-parameter turning for each dropout layer in the network. This can quickly become unwieldy. As mentioned in Section 1.1.3 and 1.1.3.1 there are many improvements proposed for standard dropout such as (Hinton et al., 2012), (Wang et al., 2019c), (Ba and Frey, 2013), (Zolna et al., 2017). However, the techniques proposed often require complex structural changes to the network, the need to train an ensemble of networks, or additional hyper-parameters. Hence, implementing these improved techniques is expensive and precludes easy adaptation into existing neural network architectures and pipelines. Data augmentation methods described in Section 1.1.3.2, as regularization techniques, are similar to dropout, and require input image changes at training time. They either remove blocks of information in the input image or add noise to the images. These methods have the drawback of requiring domain knowledge, so that the right information is dropped or added appropriately.

1.5.3 Inadequate Alignment of Source and Target Domain

A classic problem faced by all deep learning models is that the model trained on a dataset from one domain will perform poorly on datasets from another domain. For example, a model trained

to detect pedestrians using a dataset captured on a sunny day will perform poorly on a dataset captured on a rainy day. The learned feature representations of the datasets captured under different conditions have different distributions and hence are not aligned. Thus, models trained in this manner are not generalized to work for any dataset. Hence, one of the obvious strategies in supervised learning is to train the model with a dataset containing representative data from all domains. However, this approach is expensive and impractical. Hence, alternative approaches have been proposed where a deep learning model is trained in a cross-domain manner, where the source domain is labeled and the target domain is unlabeled. In this training method, we say that the source domain is supervised, and the target domain is unsupervised. There are lots of methods [(Pham et al., 2021), (Inoue et al., 2018), (RoyChowdhury et al., 2019), (Ganin et al., 2016), (Saito et al., 2019), (Sindagi et al., 2020), (Rodriguez and Mikolajczyk, 2019)] to achieve domain invariance, which has been discussed in Section 1.2.2. Broadly, the methods can be classified as, GAN-based methods, feature transfer methods, and methods with intermediate styles.

However, none of these methods provide domain alignment of features for small objects with different backgrounds and with different morphological characteristics, in applications such as mitotic cells, or in general medical applications involving cell counting. Furthermore, the methods perform poorly in applications including person ReID, which have challenges such as different camera configurations, different poses, and occlusions. The existing methods also perform poorly when the domain gap is large.

1.5.4 Weak and Incomplete Dataset Annotation

Annotating data is an expensive and time-consuming operation and is also error prone. Hence, there are a growing number of methods to learn models from un-annotated data, by transferring knowledge from annotated data. The unlabeled data could be incomplete, inexact or inaccurate, which means the data may not be labeled, partially labeled or incorrectly labeled, respectively. In the context of cell counting, weakly annotated data, is cell data labeled with dot annotations, only, rather than with bounding boxes or instance-level, per-pixel annotations. Datasets where labels are not used in cross-domain scenarios, are incompletely labeled datasets. Though there are many methods, as discussed above in the Section 1.2.1, to handle weakly supervised data, it still has challenges to overcome. In inexact annotations, including dot annotations, the dot may

appear anywhere in the cell. This poses a challenge in accurate localization of the cells. There may be a large difference in the distributions of labeled and unlabeled data, or the distributions may not adhere to model or domain assumptions. This makes the domain knowledge transfer from source domain to a target domain, a challenge. This results in the model not generalizing well on the unlabeled target dataset. Additionally, data with noisy background and overlapping objects, also pose a challenge in training a robust model.

1.6 Thesis Contributions and Organization

Despite the rapid advancement in deep learning algorithms, there are still a lot of challenges for visual content analysis tasks, as identified in previous sections. We consider these challenges in the context of their applications and propose novel solutions to address them. To address the problem of imprecise localization, we propose a framework to improve localization by automatically predicting masks and bounding boxes. To address the issue of lack of generalization techniques, we introduce a method of automatic and adaptive dropout which is both hyperparameter free, and free of any structural changes to the network. In our proposed method, the dropout rate is automatically determined and also made adaptive, during the training phase. We address the challenge of domain gap by transferring knowledge from the source domain to the target domain in two different ways. In the first instance, this is achieved by introducing a domain adaptive component with a gradient reversal layer to learn domain-invariant features at both instance and image level. In the second instance, this is achieved with self-supervised training using robust pseudo-labels generated by training a ReID network with region-specific augmentation. In both these techniques, knowledge from the source domain is transferred to the target domain where the labels are unavailable.

The localisation and domain adaptive approaches also address the lack of fully annotated datasets, as the target dataset is unlabeled. Our improved localisation methods automatically generate masks for datasets with weak annotations, thus improving overall accuracy. For the application of cell counting, we use incomplete datasets and weak annotations in the form of dots to label entire cells. We automatically generate good exemplars for such datasets using cross-domain training, which is then used to train a network to count cells by predicting a density map.

In this thesis, we have introduced two parts to group the solutions into fully supervised and weakly supervised advanced learning-based methods to handle the challenges described above. These solutions are provided in the context of various applications to provide better accuracy for detection, classification, counting and identification. In the first part of our thesis, we introduce sophisticated deep learning approaches for computer vision applicable within a fully supervised context. We present methodologies aimed at enhancing model generalization and refining object localization. We have explored the use of the advanced deep learning methods in classification and object detection to demonstrate the effectiveness of our algorithms. The fully supervised approaches provide promising results, but rely on fully annotated datasets. In the second part of our thesis, we extend our fully supervised approaches to handle incomplete datasets where the major challenges are domain gap and weak annotations, and apply it to object detection, person re-identification, and cell counting scenarios.

1.6.1 Advanced Learning-based Fully Supervised Methods

In Chapter 2, we address the challenge of overfitting and model generalization, described in Section 1.5.2, to achieve better generalization using an automatic and adaptive dropout method. We propose two methods to address this issue. In the first method, we automatically determine the dropout rate for each layer, from the distribution of the activations from the previous layer. We further introduce an adaptive method where we cluster activations from the previous layer, and for each cluster, apply different dropout rates, adaptively. Our paper (Dodballapur et al., 2020a) describes this method, which eliminates the need to tune dropout rates while also avoiding structural changes to the networks.

In Chapter 3, we tackle the challenge of imprecise localization, described in Section 1.5.1, with a multi-task cascade prediction network architecture. In our architecture, we introduce a cascade method, using region proposals. In the first step, we use Mask R-CNN, a region proposal network, to predict masks and tight bounding boxes for the cells. The prediction of masks with the bounding boxes helps to provide better localization of the cells. Thus, the localized cells, obtained, help provide better quality features. We further exploit the improved localisation by obtaining hand-crafted features, which aid in capturing morphological and textural features. The hand-crafted features are used with deep learned features, which together, provide spatial features. We extend the architecture to predict pixel-level masks for datasets which have weak

bounding-box annotations. In the second stage, we use a deep CNN network to reduce the false positives and improve the accuracy of detection. The content of this chapter is based on the paper (Dodballapur et al., 2019).

1.6.2 Advanced Learning-based Weakly Supervised Methods

In Chapter 4, we introduce an approach to address the issue of the domain gap between the source and the target domain data as described in Section 1.5.3. In particular, in this chapter, we propose a method applicable to histopathology images, where different stains and tissue preparation, result in variations in different datasets. In this method, we use a dual-stage, cascade, domain-adaptive network for mitosis cell detection to address the issue of domain gap. In both stages, a gradient reversal layer in the domain adaptation component, is used to bridge the domain gap. This method is an extension of Chapter 3, to provide an integrated, domain adaptive and better localization solution. In the first stage, we provide instance-level domain adaptation. In the second stage, we provide image-level domain adaptation. The first stage uses Mask R-CNN region proposal network to generate masks and bounding boxes for better localisation. In the second stage, false positives are filtered out to detect the mitotic cells. Additionally, in this approach, only the source dataset is labeled, while the target dataset is unlabeled. This addresses the issue described in Section 1.5.4. As Mask R-CNN is used in the first stage, the network helps provide better localisation to solve the problem stated in Section 1.5.1. The content of this chapter is based on the paper (Dodballapur et al., 2020b).

In Chapter 5, we introduce an approach to the automatic generation of good exemplars for cell counting, building on the fully-supervised localization approach introduced in first part. This approach addresses the problem of incomplete or weak annotation, discussed in Section 1.5.4, which, in turn, gives poor quality exemplars. Cross-domain exemplars are harnessed by taking advantage of the structural similarity of the cells between different domains. In this approach, we exploit the cell structure similarity, to achieve cross-domain knowledge transfer. The architecture comprises two domain-adaptive Faster R-CNN networks trained on RGB microscopy images and background-subtracted images. Based on the merged-confidence scores from these networks, the exemplars are selected and clustered. Good quality exemplars are then selected from the clusters, based on their appearance variability. A density map-based counting method is then used to count the cells based on a fraction of the selected exemplars. By choosing good

quality cross-domain exemplars, the method is able to obtain improved counting accuracy with a limited dataset, when compared to state-of-the-art methods. The content of this chapter is not available online as the corresponding paper has been submitted to International Symposium of Biomedical Imaging 2024³, and is currently under review.

In Chapter 6, we continue with the theme of domain adaptation and propose an approach to address the issue of domain gap for person re-identification as described in Section 1.5.3. In particular, in this chapter, we address this issue for person re-identification, where the same person will need to be identified across different cameras. We propose a ReID learning framework, where we use guided feature learning using region-specific augmentation. Additionally, we use a self-supervised technique to generate pseudo-labels, obtained from region-specific features, combined both temporally and spatially. This pseudo-labeled, self-supervised method helps to align the features of the source and target domains. In our case, the spatio-temporal, region-specific features help to align the region features better, leading to less label noise. Furthermore, by training with labeled and pseudo-labeled data, domain invariant features are more readily learnt. This solution also addresses the problem of incomplete or unavailable annotated datasets, as mentioned in Section 1.5.4. The content of this chapter is not available online, as the corresponding paper has been submitted to IEEE International Conference on Multimedia and Expo⁴, and is currently under review.

1.6.3 Conclusions and Future Work

In Chapter 7, we summarize our contributions and propose future directions for further study. Our contributions are broadly applicable to the domain of deep learning with a focus on visual content analysis tasks. In this thesis, we have explored broad topics such as model generalization techniques and domain adaptation both of which have wide applicability in computer vision. Additionally, we have made contributions to practical applications such as person re-identification and medical image analysis. For future study, we have identified topics related to incremental improvements over our current contributions as well as more green-field topics by drawing upon other emerging areas within the field of machine learning.

³ISBI [https : //biomedicalimaging.org/2024/](https://biomedicalimaging.org/2024/)

⁴ICME-2024 [https : //2024.ieeeicme.org](https://2024.ieeeicme.org)

Part 1

Advanced Learning-based Fully Supervised Methods

Automatic Dropout for Deep Neural Networks

Neural networks typically have a large number of parameters, which leads to overfitting. Hence, the trained models do not generalize well, and perform poorly on unseen data. Dropout is a well understood regularization technique used for reducing co-dependence of the neurons leading to a more generalized deep learning model. Standard dropout requires hyper-parameter tuning for each dropout layer in the network which can become unmanageable as the layers increase. As described in Section 1.1.3 several improvements have been studied for standard dropout which require complex structural changes to the network, training ensemble of networks, additional hyper-parameters, or domain knowledge for augmentation. These techniques are complex, expensive and precludes easy adaptation into existing neural network architectures and pipelines. In this chapter, we introduce a method of sampling a dropout rate from an automatically determined distribution. We further build on this automatic selection of dropout rate, by clustering the activations, and adaptively applying different rates to each cluster. We show that our methods outperform standard dropout across all datasets and neural networks. The content of this chapter is related to the publication (Dodballapur et al., 2020a).

2.1 Introduction

Deeper and wider convolutional neural networks (CNNs) have been gaining popularity over the last few years in the quest for better feature representation. This has resulted in neural networks with a greater number of network parameters. It is well known that such networks are typically prone to overfitting. Hence, in practice, regularization methods have been proposed to improve generalization, and ultimately network performance. A number of regularization methods have been proposed, such as L2 regularization, early stopping, data augmentation, and dropout, to achieve better model generalization, and prevent model overfitting.

Of all the regularization techniques, the standard dropout (Hinton et al., 2012) and its proposed improvements (Labach et al., 2019) have been prolific in the advancement of neural network performance since the application of standard dropout in 2012 to win the Large Scale Visual Recognition Challenge (Krizhevsky et al., 2012). The popularity of dropout has also been helped by advances in the theoretical understanding of why dropout works. According to the seminal paper (Hinton et al., 2012), dropout is one of the regularization methods, which works by reducing co-adaptation of the neurons.

A limitation of the dropout method is that it typically requires a dropout rate hyper-parameter, which defines the probability of a neuron being dropped or retained during training. This parameter, if set too high, results in many neurons being dropped, affecting the convergence of training, and if set too low, does not generalize the model well. Additionally, most of the deep network implementations have a fixed dropout rate for all layers of the network. In the original dropout implementation (Hinton et al., 2012), the dropout rate was set to 0.5. Usually, the optimal dropout rate for a network is determined through cross-validation. However, determining an optimal dropout rate for each dropout layer of the network quickly becomes unwieldy. The following section summarizes the related work in this field.

2.1.1 Related Work

There have been several approaches (Keshari et al., 2019), (Singh et al., 2016), (Wang et al., 2019c) proposed in the literature, which improve on standard dropout, by focusing on improving regularization, and rate of convergence during training (Tompson et al., 2015), (Wang and Manning, 2013). The proposed methods are broadly categorized into adaptive and stochastic dropout approaches (Labach et al., 2019). Whilst some adaptive dropout techniques are based on structural changes to the network (Keshari et al., 2019), other adaptive techniques have explored the use additional networks (Zolna et al., 2017) during training. The Gaussian dropout method (Srivastava et al., 2014) is a stochastic technique which achieves the dropout effect by adding a predetermined Gaussian noise to the neurons, rather than dropping the neurons. A further extension of this method, automatically learns the parameters of a Gaussian noise distribution, as part of the network training process, by learning a variational objective (Kingma et al., 2015). This, in essence, leads to learning a dropout rate for each layer of the network. A more recent stochastic approach, named jumpout (Wang et al., 2019c), proposes a method of

sampling the dropout rate from a Gaussian distribution with a pre-determined standard deviation.

Fast dropout (Wang and Manning, 2013) proposes a method for faster training convergence, by showing that activations during training with dropout, can be approximated as a Gaussian distribution. Variational dropout (Zhuo et al., 2015) extends Fast dropout by introducing a Bayesian feature noising model to infer dropout noise for neurons. Targeted dropout (Gomez et al., 2019) proposes a method of dropping a certain proportion of low-valued weights, to reduce degradation of performance due to dropout, towards the end of training. Curriculum dropout (Morerio et al., 2017) proposes a method of varying dropout, to make learning more difficult as the network training progresses.

Rather than dropping neurons from a specific dropout layer, Stochastic Depth (Huang et al., 2016) proposes dropping entire network layer of a deep neural network during training, while retaining the entire network depth during testing. SwapOut (Singh et al., 2016) generalizes on both dropout and stochastic depth, where activations are randomly assigned the values zero, input value, output value, or sum of input and output values. There is also a family of adaptive dropout techniques that rely on additional neural networks or neural network elements. Standout (Ba and Frey, 2013) is an adaptive method, where a deep belief network and a deep neural network (DNN) with shared parameters, are trained jointly, to obtain the dropout rate adaptively across all layers. In Fraternal dropout (Zolna et al., 2017), the network is trained to be invariant to dropout, by training two identical networks with shared parameters, but with different dropout masks.

2.1.2 Contributions

In this chapter, we tackle the challenge of overfitting and poor model generalization, outlined in Section 1.5.2. Our proposed advanced methods overcomes the limitations of changes suggested in Section 1.1.3 which require complex structural changes to the network, the need to train an ensemble of networks, additional hyper-parameters, or require domain knowledge for augmentation. The effectiveness of our methods, though demonstrated for object classification, in a fully supervised setting, is applicable to any deep neural network. Our proposed two methods for improving the performance of the neural networks using automatic and adaptive dropout are as follows:

- In our first method, we automatically determine the dropout rate to be used for a layer, by using the distribution of neuron activations from the previous layer;
- In our second method, we cluster the activations using K -means, and apply different dropout rates to activations in each cluster.

Our methods are widely applicable to a variety of neural networks such as autoencoders, deep neural networks for classification and segmentation, where dropout is used for model regularization. Our methods do not rely on making structural changes to the network or additional network elements during training. Unlike the standard dropout, we eliminate the need to search the hyper-parameter space, for an optimal dropout rate. Compared to the most recent development, jumpout (Wang et al., 2019c), our method differs by estimating the parameters of the Gaussian automatically from the activations of each network layer.

Standard dropout (Hinton et al., 2012) requires a hyper-parameter for every layer with dropout function enabled but our method removes the need for this hyper-parameter. Unlike methods such as (Ba and Frey, 2013) and (Zolna et al., 2017), our method does not introduce any structural changes to the network, or need an additional network during training. Unlike DropConnect (Wan et al., 2013), our method does not apply sparsity to weights and biases of neurons. Furthermore, unlike Bayesian methods such as the Fast dropout (Wang and Manning, 2013) and Variational dropout (Kingma et al., 2015), our method does not involve modelling of noise for each neuron. Finally, all of the above methods need a hyper-parameter, which is either in the form of a dropout rate, or parameters of a Gaussian for their respective implementation.

Our automatic dropout rate method is inspired by jumpout (Wang et al., 2019c), in that we too sample a dropout rate from a Gaussian distribution but unlike jumpout, which requires cross-validation to establish the parameters of the Gaussian distribution, our proposed automatic dropout rate selection method determines it from the activation of the neurons at each individual layer. Our second method that adapts dropout rates to clusters of activations is inspired by a recent method called Guided dropout (Keshari et al., 2019). Our method, instead of learning the strength of a neuron as they do in (Keshari et al., 2019), applies clustering to the activations at training time, and applies different dropout rates to each cluster. In our implementation, we modify the dropout rate determined automatically, in accordance with the cluster centre of each cluster.

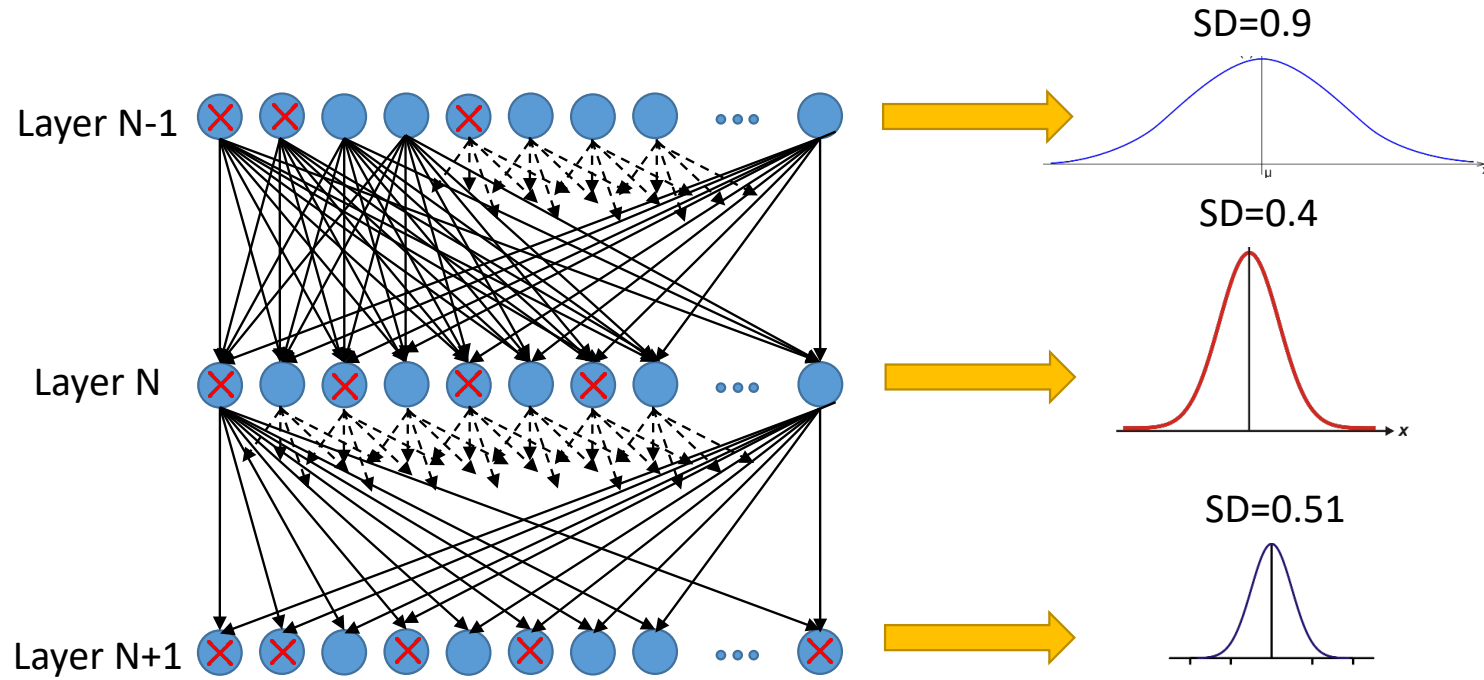


FIGURE 2.1: Automatic dropout rate determination at the three layers of a neural network. The distribution of neuron activations is approximated as a Gaussian with the determined standard deviation. The dropout rate for each layer is determined from this Gaussian distribution. The dropped neurons are shown with a red cross against them. The standard deviation of activations at layer $N - 1$ is determined as 0.9, and for layer N , the dropout rate is sampled from a Gaussian distribution with $SD = 0.9$. Similarly, for layer $N + 1$, the dropout rate is sampled from Gaussian with $SD = 0.4$.

2.2 Method

We propose two methods of choosing a dropout rate automatically, eliminating the need to choose a hyper-parameter for dropout for every given layer of the network. In our first method, we choose a dropout rate for a layer by analysing the distribution of neuron activations from the previous layer. We further enhance this method by clustering activations and applying an adaptive dropout for the activations based on their membership to a cluster. Both these methods are described in more detail in the following sections.

2.2.1 Automatic Dropout

As observed by existing studies (Wang et al., 2019b), (Wang et al., 2016b) and (Raghu et al., 2017), a DNN which uses ReLU activations can be thought of as piecewise linear weight function which divides the input plane into regions. For non-zero bias terms which are usually the case with neural networks, the regions form a convex polyhedra. The shape of the polyhedra is based on the activation patterns, each represented by a set of linear constraints.

Our automatic dropout rate method relies on finding smooth approximations to such convex polyhedra. Since ReLU is a cheap and efficient activation function used in state-of-the-art deep networks, our research is focused on improving the performance of such deep networks. Dropping out some activations changes the polyhedral accordingly. Thus, the dropout mechanism trains a family of DNNs, each represented by a slightly different polyhedron. A constant dropout rate can be considered as the average distance between the polyhedra created by dropout. So, the weights of the final model can be thought of as an average of these individual polyhedral regions created as a result of dropout. To obtain good generalization performance, the resulting averaging should result in locally smooth regions. However, with fixed dropout, it can only be achieved when the dropout rate is low.

As in jumpout (Wang et al., 2019c), we sample the dropout rate from a Gaussian distribution which ensures that the probability of dropping out more activations reduces as the number of units dropped out increases. We further extend the work of jumpout (Wang

et al., 2019c) by automatically choosing a Gaussian distribution based on the spread of the activated neurons. This ensures that the probability of obtaining a smooth polyhedra not only increases but it is also locally optimized by sampling from a Gaussian distribution of an appropriate shape.

In our method, we automatically determine the dropout rate for a dropout layer based on the standard deviation of the distribution of activations of the neurons from the earlier layer. We then sample a dropout rate from this distribution and determine a dropout mask based on the dropout rate. To ensure that during the training phase neither too many activations are dropped out nor too few activations are dropped out, we further truncate $|p|$ so that $|p| \in [s_{min}, s_{max}]$, where $0 \leq s_{min} < s_{max} \leq 1$. We note that the standard deviation σ will have a lower value if the activations are all similar to one another when compared to the case where there is a large variation in the magnitude of the activations.

The details of the automatic dropout rate method are illustrated in Algorithm 1. In our implementation, this algorithm is implemented in a separate custom dropout layer into which the activations a_j from the previous j^{th} layer are input. Hyper-parameters s_{min} , s_{max} are set such that the sampled dropout rate is clipped to be within acceptable bounds.

Algorithm 1 Automatic Dropout (M_1).

Input a_j, s_{min}, s_{max} where a_j are activations of the j^{th} layer.

Compute the mean of the activations.

$$a_{ave} = (\sum a_j) / n$$

Compute standard deviation of activations.

$$a_{std} = \sqrt{(\sum (a_j - a_{ave})^2) / (n - 1)}$$

Compute dropout probability.

$$N(0, a_{std}), p_{trunc} = \min(s_{min} + p, s_{max})$$

Apply dropout to activations with a rate p_{trunc} .

The details of automatic dropout rate algorithm are also illustrated in Figure 2.1, where a part of a neural network with 3 layers, $N-1$, N and $N + 1$ is shown. For each layer, during training, we determine the standard deviation of the activations. In this example,

for layer $N-1$ the computed standard deviation is 0.9, for layer N it is 0.4 and for layer $N+1$ it is 0.51. The computed standard deviation is used to determine the parameters of the Gaussian distribution to sample a dropout rate from. In all our experimental results, we refer to this method by name as M_1 .

2.2.2 Adaptive Dropout

Inspired by (Keshari et al., 2019), we hypothesize that varying dropout rates based on the activation itself would allow the network to generalize better. Thus, we propose an enhancement over standard dropout to adaptively dropout neurons based on the strength of the activation. Our adaptive dropout strategy is based around the clustering of activations and applying varying levels of dropout rates for each cluster. The details of the adaptive dropout based on clustering are illustrated in Algorithm 2. In our implementation, this algorithm is implemented in a separate custom dropout layer along with the automatic dropout method to which the activations a_j from the previous j^{th} layer are input. Hyperparameters s_{min} , s_{max} are set to values so that the sampled dropout rate is clipped to be within acceptable bounds.

Algorithm 2 Adaptive dropout using clustering (M_2).

Input a_j, s_{min}, s_{max} where a_j are activations of the j^{th} layer.

Compute the mean of the activations.

$$c_k = kmeans(a_j)$$

for $k = 1$ to 3 **do**

 Determine the cluster with the highest mean.

$$c_{max} = max(c_k)$$

end for

Compute p_{trunc} according to either algorithm M_1 or use a fixed dropout rate.

for $k = 1$ to 3 **do**

 Compute dropout probability for each cluster k .

$$p_k = p_{trunc} * (c_k / c_{max})$$

end for

Apply dropout p_k to activations belonging to cluster k .

In (Keshari et al., 2019), a separately learned parameter called strength of a node, a scalar value indicating the importance of the node/neuron, is learned during training. In our method, we simplify this formulation by using the magnitude of the activation as a proxy for the strength. Further, we perform K -means clustering of the activations at training time and group the activations into N clusters. We then order the clusters in descending order of the cluster mean, and apply a dropout rate, which is inversely proportional to the cluster mean (details later in this section). The intuition underlying this dropout strategy is as follows. We want to progressively decrease the dropout rate as the strength of activation decreases. Similar to (Keshari et al., 2019), we argue that by dropping higher strength activations at a higher dropout rate, the network generalizes better by relying on activations of a lower strength. Note that this method can use a fixed dropout rate (dropout as a hyper-parameter), as well as the automatically determined dropout rate proposed earlier, and progressively adapt to different clusters of activations.

We rely on the study (Mianjy et al., 2018) to provide a theoretical understanding of why the adaptive dropout using clustering works. The process of dropout as a regularization method can be understood by considering a neural network with input $x \in \mathbb{R}^{d_2}$. Let $y \in \mathbb{R}^{d_1}$ be the output feature vector obtained by the neural network represented by the matrix M . Then, $y = Mx$ for some $M \in \mathbb{R}^{d_1 \times d_2}$. Now, if we assume that the neural network has two hidden layers with respective weight matrices $U \in \mathbb{R}^{d_1 \times r}$ and $V \in \mathbb{R}^{d_2 \times r}$, the goal of learning is to determine the weight matrices corresponding to the hidden layers U, V such that the following objective in Eq. (2.1) is minimized:

$$f(U, V) = \mathbb{E}[\| (y - UV^T x) \|^2]. \quad (2.1)$$

In the presence of dropout, Eq. (2.1) can be rewritten as follows:

$$f(U, V) = \mathbb{E}[\| (y - \frac{1}{\theta} U \text{diag}(b) V^T x) \|^2], \quad (2.2)$$

where $b \in \text{Bernoulli}(\theta)$, which means the indicator function b is sampled from Bernoulli trials. In (Mianjy et al., 2018), Lemma A.1 is used to show that Eq. (2.2) can be rewritten

as follows:

$$f(U, V) = l(U, V) + \lambda \sum \|u_i\|^2 \|v_i\|^2, \quad (2.3)$$

where u_i, v_i are the columns of the matrices U, V respectively and λ is a regularization constant equal to $\frac{1-\theta}{\theta}$. The term $\lambda \sum \|u_i\|^2 \|v_i\|^2$ increases in value, hence increasing the overall loss $f(U, V)$ when higher activations are dropped. This allows the network to rely on activations of a lower magnitude to improve their contribution to the overall network generalization. A similar principle is used in (Keshari et al., 2019) with the strength parameter.

The details of the adaptive dropout based on the clustering algorithm are illustrated in Figure 2.2. A key feature of this method is to use the magnitude of a neuron activation as a guide to regularizing the network. We argue that just as using large weights leads to network overfitting, large activations also contribute to overfitting. The adaptive dropout through clustering addresses this issue.

As illustrated in Figure 2.2, we cluster the activations for each layer independently. In our implementation, we use three clusters which map to low, medium, and high strength activations. We adaptively apply the dropout rate to each cluster as follows. For the cluster with the largest centroid, we apply the dropout rate determined according to either Algorithm 1 or using fixed dropout rate. For the next highest cluster, we reduce the dropout rate by an amount c_k/c_{max} where c_k is the cluster mean and c_{max} is the maximum cluster mean. This process is repeated for all the clusters. This allows the network to not overly rely on higher strength activations during training and leads to better generalization. In all our experimental results, we refer to this method by name as M_2 . For the ablation study, we also design a variation of this method, where we adaptively apply the dropout rate with the highest dropout rate assigned to the cluster with the lowest centroid. In all our experimental results, we refer to this method as M_2^{min} . When this method is evaluated independently of the automatic dropout rate method, we use a fixed dropout rate as input.

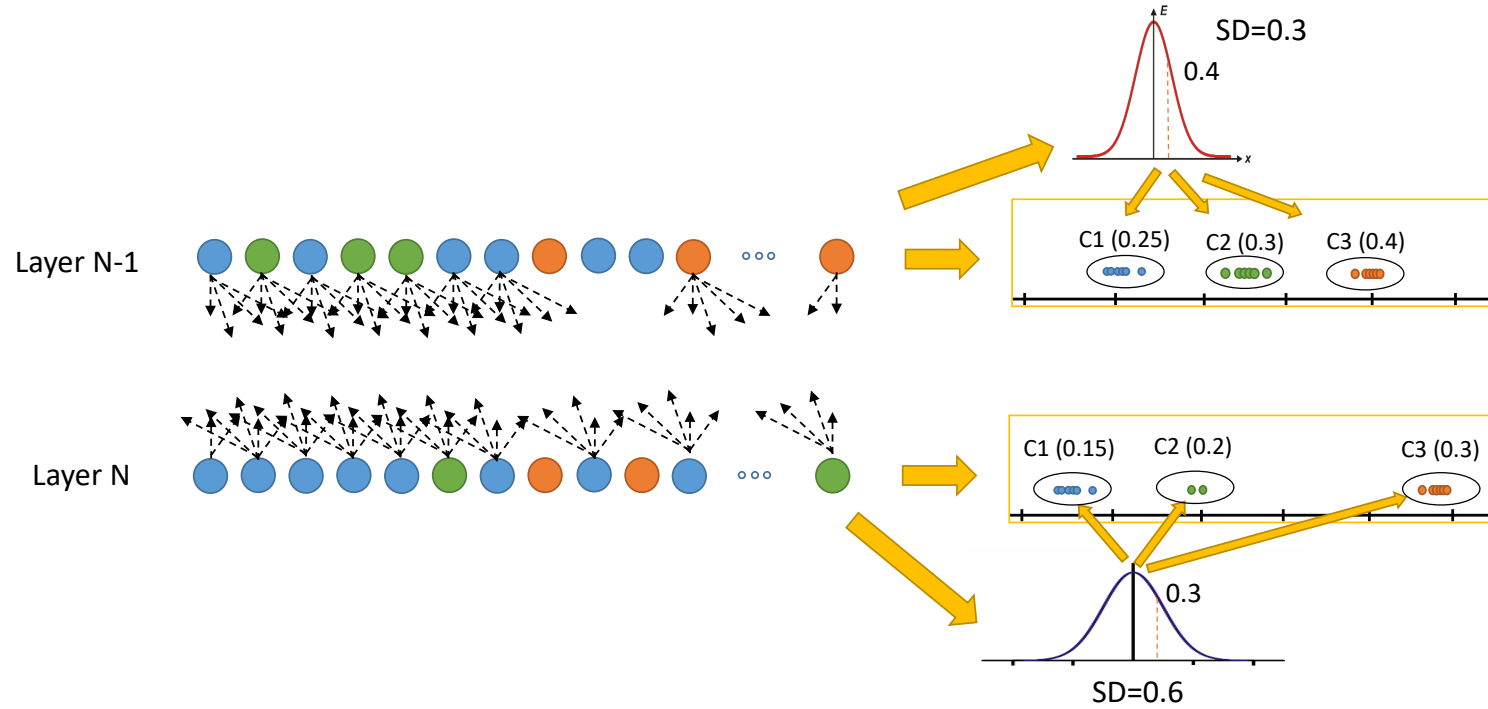


FIGURE 2.2: Adaptive dropout based on clustering working in combination with automatic dropout rate selection. The clustering of activations allows different dropout rates to be applied selectively to activations within a cluster, based on the strength of the activation. At each layer, as illustrated, three clusters are determined, based on the neuron activations. Different dropout rates are applied to each cluster, based on the strength of the activation of each cluster. For determining the strength of activation, the cluster centroid's activation is used. The adaptive dropout rate applied is derived from an empirically determined fixed dropout rate or from a distribution of activations of the previous layer. As an example, for layer N , there are three clusters with centroids 0.15, 0.2 and 0.3, and the dropout rate of 0.3 is adaptively applied to each cluster. This rate is sampled from a Gaussian of $SD = 0.6$.

2.3 Experiments and Results

2.3.1 Datasets

We chose Fashion-MNIST (Xiao et al., 2017), CIFAR-10 (Krizhevsky et al., 2009), and CIFAR-100 (Krizhevsky et al., 2009) datasets, which are image classification datasets, for our experiments. Sample images from CIFAR-10 and Fashion-MNIST datasets are shown in Figure 2.3.

CIFAR-10: This is an image classification dataset consisting of 60,000 tiny RGB images. Each image is of size 32×32 . There are 50,000 images in the training set and 10,000 images in the testing set. There are 10 mutually exclusive classes in the dataset. Hence, in the training set, there are exactly 5000 image per class. Similarly, in the test set there are 1000 images per class.

CIFAR-100: Similar to CIFAR-10, this is also an image classification dataset. In CIFAR-100, there are 100 classes which are used as fine labels of the images. These 100 classes belong to 20 super classes which are used to specify the coarse labels of an image. CIFAR-100 dataset consists of 60,000 tiny RGB images, where each image is of size 32×32 . There are 50,000 images in the training set and 10,000 images in the testing set. Thus, in the training set, there are 500 images per fine label class, and in the test set, 100 images per fine label class.

Fashion-MNIST: This dataset consists of 70,000 tiny grey scale images from the Zalando's article image collection¹. There are 60,000 images in the training set and the rest in the testing set. Fashion-MNIST is a drop-in replacement for the MNIST² dataset and has exactly 10 classes. The image size is 28×28 .

¹Zalando collection <https://jobs.zalando.com/en/tech/>

²MNIST Dataset <https://yann.lecun.com/exdb/mnist/>

2.3.2 Implementation and Results

To measure the improvement provided by using dropout and its variants, we kept data augmentation to a minimum. We ran the evaluation of each method for each dataset and network 3 times and reported the median result in each case. All our experiments were performed using two NVIDIA GeForce GTX 1080 Ti GPUs and the Keras/TensorFlow programming platform. We used Wide residual network (Zagoruyko and Komodakis, 2016) (Wide ResNet) and v3 network³ as the baseline. We chose Wide ResNet (Zagoruyko and Komodakis, 2016) because it has demonstrated state-of-the-art results over other popular networks. The variants of Wide ResNet, namely $width = 8$, $depth = 16$ and $width = 10$, $depth = 28$ of Wide ResNet, have fewer parameters than ResNet but have performed better on CIFAR (Krizhevsky et al., 2009). In our experiments involving Wide ResNet and CIFAR datasets, we used stochastic gradient descent (SGD) with Nesterov momentum and cross-entropy loss. The initial learning rate was set to 0.1, weight decay to 0.0005, dampening to 0, and momentum to 0.9. We trained the network for 200 epochs, where the learning rate was dropped by 0.2 at 60, 120, and 160 epochs. In our experiments involving the Fashion-MNIST dataset and Wide ResNet, the initial learning rate was set to 0.001, weight decay to 0.0005, dampening to 0, and momentum to 0.9. We used a time-based learning decay of $1e-5$. We trained the network for 50 epochs. For both CIFAR and Fashion-MNIST experiments, mini-batch size was set to 64, and horizontal flip was the only data augmentation used. The training protocol was very similar to (Zagoruyko and Komodakis, 2016).

We also performed experiments on a simpler network similar to VGG called the v3 network. This was to apply our methods to a very different architecture when compared to Wide ResNet. For our experiments with the v3 network, we used SGD with an initial learning rate of 0.01 and momentum of 0.9. We trained the network for 50 epochs, where we dropped the learning rate by 0.5 and used momentum decay of 0.05 every 10 epochs. This training protocol is similar to the implementation of v3 network.

For standard dropout, we empirically determined the best performing dropout ($= 0.3$) for Wide ResNet across all datasets, and used this value for all the layers of Wide ResNet

³The “v3” network from https://github.com/jseppanen/cifa_lasagne

across all our experiments. For the proposed automatic dropout rate, we limited the automatically determined dropout rate such that it always lay in the interval $[0.01, 0.6]$. We compared our implementation with the standard dropout (Hinton et al., 2012) as well as jumpout (Wang et al., 2019c). For the v3 network, we have compared our method with the standard dropout only, since the jumpout method did not converge for the datasets we experimented with.

2.3.3 Ablation Experiments

To get a better understanding of our proposed methods, we performed several ablation experiments by combining our two proposed methods and introducing variations on our proposed methods. For these ablation experiments, we used the three variants of Wide ResNet, namely, $width = 8, depth = 16$, $width = 10, depth = 28$, and $width = 4, depth = 40$. We chose these variants because they explore the network architecture both in terms of width and depth. We considered a simple training protocol with no data augmentation, an initial learning rate of 0.01, SGD with Nesterov momentum, cross-entropy loss, and time-based learning rate decay of $1e-5$. We trained for 50 epochs. We chose to use no data augmentation to measure the regularization effect of our methods alone.

Table 2.1 shows the classification accuracy results of our method using the CIFAR-10, CIFAR-100, and Fashion-MNIST datasets for the two variants of the Wide ResNet we evaluated on. We used the same augmentation method and training protocol (Zagoruyko and Komodakis, 2016) for all methods and hence can compare the relative performance of our proposed methods against both standard dropout and jumpout. We note that across both variants of Wide ResNet networks tested, our methods outperform the standard dropout and jumpout for the CIFAR-10, CIFAR-100, and Fashion-MNIST using the automatic dropout rate method. The combination of the first and second methods gave a small improvement over the first method alone.

Table 2.2 shows the classification accuracy results on the v3 network. For the v3 network, we ran experiments for dropout and our proposed methods only. We note that, as in the case of Wide ResNet, our automatic dropout rate method, as well as the combination

TABLE 2.1: Results comparing our method with the state-of-the-art method jumpout, and the standard dropout, for different variants of Wide ResNet ($width = 18, depth = 8$ and ($width = 28, depth = 10$) and popular datasets. CIF10, CIF100 and FASHN refer to the CIFAR-10, CIFAR-100, Fashion-MNIST datasets, respectively. The result for M_1 outperforms the state-of-the-art for both Wide ResNet variants, and $M_1 + M_2$ outperforms M_1 and state-of-the-art.

Method	Wide ResNet:16_8			Wide ResNet:28_10		
	CIF10	CIF100	FASHN	CIF10	CIF100	FASHN
Dropout	0.945	0.762	0.911	0.949	0.776	0.910
Jumpout	0.939	0.752	0.912	0.926	0.735	0.921
M_1	0.949	0.769	0.925	0.953	0.780	0.921
$M_1 + M_2$	0.951	0.771	0.931	0.956	0.785	0.932

TABLE 2.2: Results our method compared with standard dropout for the v3 network and popular datasets. CIF10, CIF100 and FASHN refer to the CIFAR-10, CIFAR-100, Fashion-MNIST datasets, respectively. The result for M_1 outperforms the standard dropout, and $M_1 + M_2$ outperforms M_1 and the standard dropout.

Method	CIF10	CIF100	FASHN
Dropout	0.845	0.481	0.920
M_1	0.874	0.495	0.923
$M_1 + M_2$	0.884	0.565	0.932

of automatic dropout rate and adaptive dropout with clustering, outperform the standard dropout.

Table 2.3 and Table 2.4 show the classification accuracy results of our ablation experiments. Firstly, we note that the results across different datasets are slightly lower than the state-of-the-art because data augmentation, and highly tuned learning rates, were not used. Here, we show that, across all networks and datasets, our first method (M_1) outperforms the standard dropout as well as jumpout (Wang et al., 2019c). Furthermore,

combining the first and second methods ($M_1 + M_2$) performs better than the first method alone. However, we note that introducing M_2 , slows down training performance by a factor of 5, when compared to M_1 alone because the K -means algorithm is computationally expensive. We hope to address this in future work. We also note that, in some instances, our second method (M_2) with a carefully chosen dropout rate performs as well as the first and second methods combined ($M_1 + M_2$). This confirms the utility of both methods in their own right. Finally, the variation on the second method as described earlier (M_2^{min}), on its own and in combination with first method (M_1) performs worse but is still comparable to jumpout for Fashion and CIFAR-10 datasets, confirming our original hypothesis about dropping high strength activations more often, for better generalization.

TABLE 2.3: Results of the ablation experiments comparing our method in different settings with jumpout and the standard dropout, using Wide ResNet variants ($width = 18, depth = 8$ and ($width = 28, depth = 10$) and popular datasets. CIF10, CIF100 and FASHN refer to the CIFAR-10, CIFAR-100, Fashion-MNIST datasets, respectively. The result for $M_1 + M_2$ or M_2 alone outperforms all other methods including state-of-the-art. Method M_2^{min} performs worse than M_2 .

Method	Wide ResNet:16_8			Wide ResNet:28_10		
	FASHN	CIF10	CIF100	FASHN	CIF10	CIF100
Dropout	0.910	0.856	0.549	0.910	0.810	0.612
Jumpout	0.910	0.834	0.548	0.920	0.820	0.623
M_1	0.925	0.874	0.632	0.920	0.850	0.636
$M_1 + M_2$	0.932	0.878	0.641	0.932	0.889	0.644
M_2	0.934	0.870	0.600	0.929	0.886	0.649
$M_1 + M_2^{min}$	0.920	0.870	0.600	0.929	0.870	0.607
M_2^{min}	0.915	0.858	0.589	0.929	0.862	0.575

TABLE 2.4: Results of the ablation experiments on Wide ResNet:40_4 ($width = 40, depth = 4$) using popular datasets, comparing our method in different settings, and comparing it with jumpout and standard dropout. CIF10, CIF100 and FASHN refer to the CIFAR-10, CIFAR-100, Fashion-MNIST datasets, respectively. The result for $M_1 + M_2$ outperforms the all other methods including state-of-the-art. Method M_2^{min} performs worse than M_2 .

Method	FASHN	CIF10	CIF100
Dropout	0.910	0.780	0.591
Jumpout	0.920	0.830	0.592
M_1	0.920	0.862	0.612
$M_1 + M_2$	0.923	0.870	0.618
M_2	0.921	0.868	0.611
$M_1 + M_2^{min}$	0.920	0.844	0.547
M_2^{min}	0.918	0.833	0.546

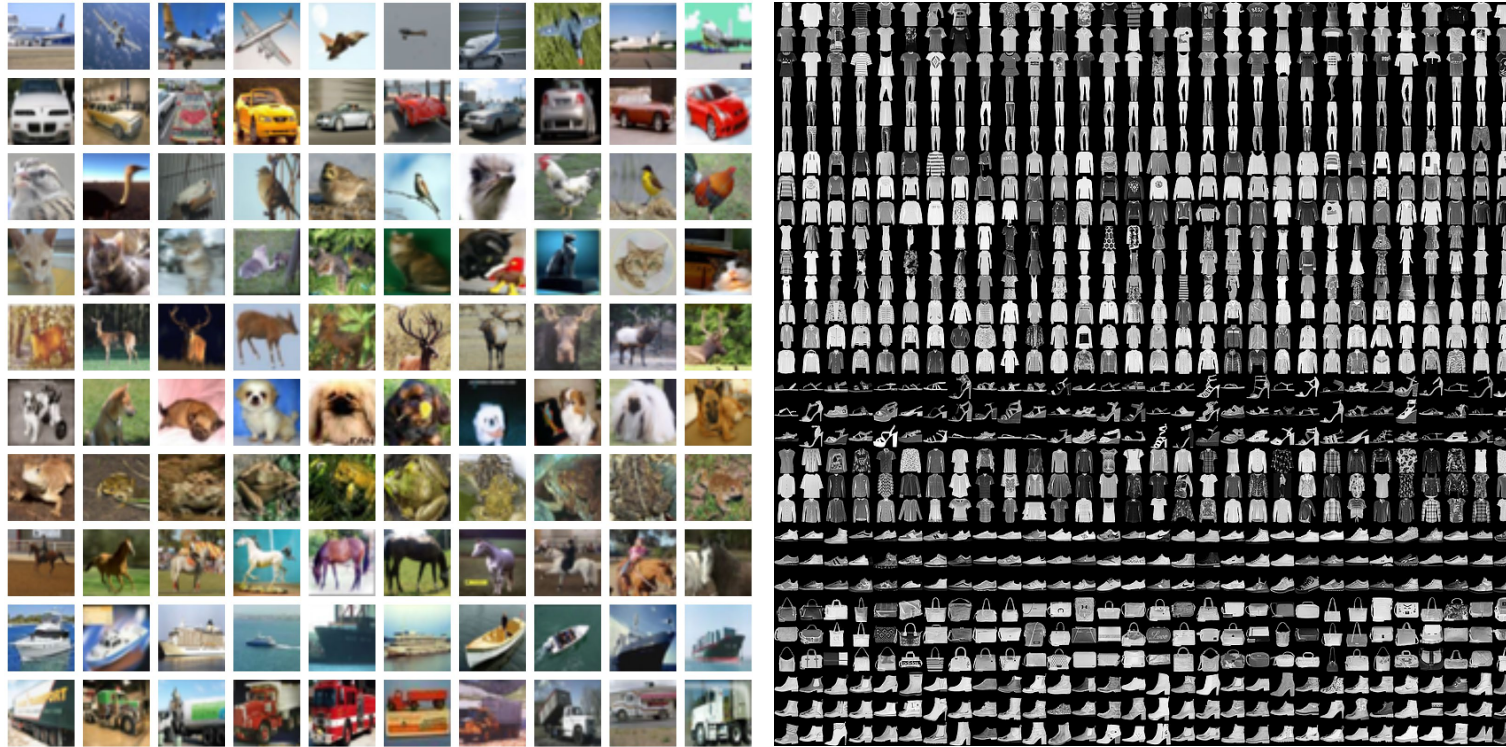


FIGURE 2.3: Sample images from (Left) CIFAR-10 dataset (Krizhevsky et al., 2009) and (Right) Fashion-MNIST dataset (Xiao et al., 2017). CIFAR-10 dataset has 10 random images per row for each of the 10 classes in the dataset. Fashion-MNIST dataset has 3 rows of images per class and has 10 classes in the dataset.

2.4 Summary

In this chapter, we have proposed two methods for improving dropout. Our methods do not require structural changes or ensembles. Our first method removes the need for choosing a dropout rate hyper-parameter. Our second method applies dropout rate adaptively to activations based on clustering. We have demonstrated that both our methods outperform standard dropout, when applied individually, as well as when combined. We have demonstrated the versatility and applicability of our methods, by evaluating on well-known datasets such as CIFAR and Fashion-MNIST using three variants of Wide ResNet, and a substantially different network such as the v3 network. We have demonstrated the effectiveness of this fundamental generalization method in this chapter. This dropout method has wide applicability across many deep neural networks, especially with many layers, where the dropout rates have to be set adaptively and automatically.

Mask-Driven Mitosis Detection in Histopathology Images

Object detection and localization methods described in Sections 1.1.2 and 1.1.1, propose bounding box methods such as Faster R-CNN (Ren et al., 2015) and instance-based methods such as Mask R-CNN (He et al., 2017). Detection approaches without masks, do not provide accurate localisation, as they do not capture the morphology of the object and may include additional noise of other objects and background, forming an inexact annotation which will have an effect on object classification accuracy, due to noisy learned feature representations. Instance-based segmentation provides an advantage with better localization, but have further challenges due to lighting changes, cluttered background, and object morphological variations. Furthermore, preparing datasets with pixel-level annotations is expensive in scenarios where there are small and overlapping objects. This chapter addresses these challenges in the context of mitosis cell detection.

The count of mitosis cells in histopathology images is an indication of the progression of cancer. However, pathologists find this task laborious, as there are a number of cells which are similar to mitosis cells. Additionally, the shape of the cells varies depending upon the phase of cell division. Several pixel-based or region-based deep learning solutions have evolved over the recent years to address this problem. Pixel-based solutions are accurate but computationally expensive. Region-based methods do not localize the cells effectively, hence having an effect on mitotic count accuracy. The solutions are designed to use both hand-crafted and learned deep features, work on pre-processed images only. In this chapter, we introduce a cascade, mask-driven approach, which helps to localize the cells better, and hence achieve a greater accuracy of mitosis cell detection. Our method predicts masks for cells automatically and uses the generated masks for mitotic detection. In the first stage of processing, we use the Mask R-CNN network to predict the masks for the cells, and also classify the cells with a high recall rate and low precision. These

cells, detected in the first stage, are processed further in a second stage, to eliminate false positives. In the second stage, the cell candidates are classified as mitotic or non-mitotic, using a combination of both hand-crafted features, and learned deep features obtained from a CNN, pre-trained on ImageNet. We have evaluated our method on the public ICPR 2012 and ICPR 2014 mitosis datasets and obtained improved performance over fully-supervised segmentation methods, as well as those using a combination of hand-crafted and learned deep features. We also show that masks learned from the small ICPR 2012 dataset can be effectively transferred to the ICPR 2014 dataset, which lacks mask annotations for cells. The content of this chapter is related to the publication (Dodballapur et al., 2019).

3.1 Introduction

Technological advances in fast-slide scanners have had a cascading effect in the health industry for improved diagnosis of diseases such as cancer (Webster and Dunstan, 2014). Cancer is one of the leading causes of death in the world today (Ferlay et al., 2015). Mitosis detection and counting are used as the primary diagnostic factors for cancer detection and grading (Veta et al., 2015). Mitosis detection is challenging because of its hyperchromaticity, and close resemblance to other cells. Additionally, it has other confounding factors, namely, sparsity, and morphological changes caused by different stages of cell division (Veta et al., 2015). Mitotic counting is a very tedious and subjective exercise, with pathologists often having differing opinions about cancer grading. Automating mitotic counting using histopathology images, can remove the subjectivity from the counting process, reduce the burden on pathologists, and facilitate a faster, more consistent, and quantitative diagnosis of cancer. The following section summarizes the related work in this field.

3.1.1 Related Work

3.1.1.1 Hand-Crafted Feature Approaches

Some of the earlier studies, have proposed automatic detection of mitotic cells using hand-crafted features, and classification methods such as, support vector machines (SVMs), decision trees, and AdaBoost (H.Irshad, 2013), (C. Sommer and Gerlich, 2012), (Tek, 2013).

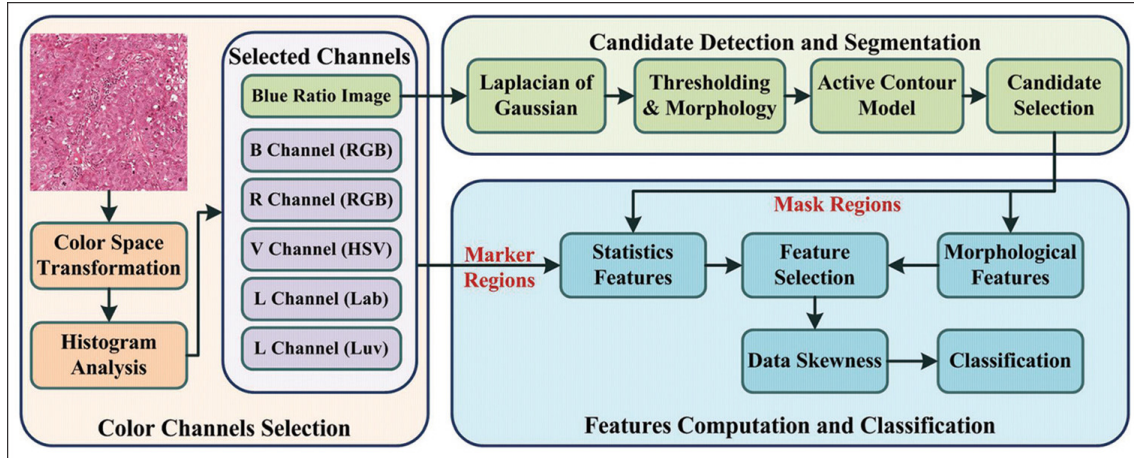


FIGURE 3.1: A three-stage process of mitosis detection. In the first stage, RGB images are transformed to other color spaces and specific color channels are selected using histogram analysis. In the second stage, a blue ratio image is determined which is useful in differentiating nuclei from background. In the third stage, morphological and statistical features are computed to train a classification model for mitosis classification (H.Irshad, 2013).

In the hand-crafted approach, the images are pre-processed, and the candidate regions or cells are identified. Then, in a feature extraction step, features such as, colour, texture, shape, and morphological features are extracted from pre-processed images and used to train a classifier (H.Irshad, 2013), (C. Sommer and Gerlich, 2012), (Tek, 2013). Morphological features are extracted from blue-ratio images, after obtaining active-contours (H.Irshad, 2013). These features, along with colour features in different spaces, are combined and used to train a decision tree model for classification of mitosis cells. Figure 3.1 provides an overview of this approach.

The study (C. Sommer and Gerlich, 2012), proposes a dual-stage mitosis cell detection approach. In the first step, a random forest classifier is trained with an interactive GUI, to segment the candidate mitosis cells using active learning. These cells are then classified using a Support Vector Machine (SVM) in the second classification step, using the texture and shape features. The study (Tek, 2013), uses RGB channels and shape features to train an AdaBoost cascade for mitosis detection. The drawback of this traditional approach is that identifying and extracting discriminative features is too time-consuming. Additionally, the method is not robust to morphological variations.

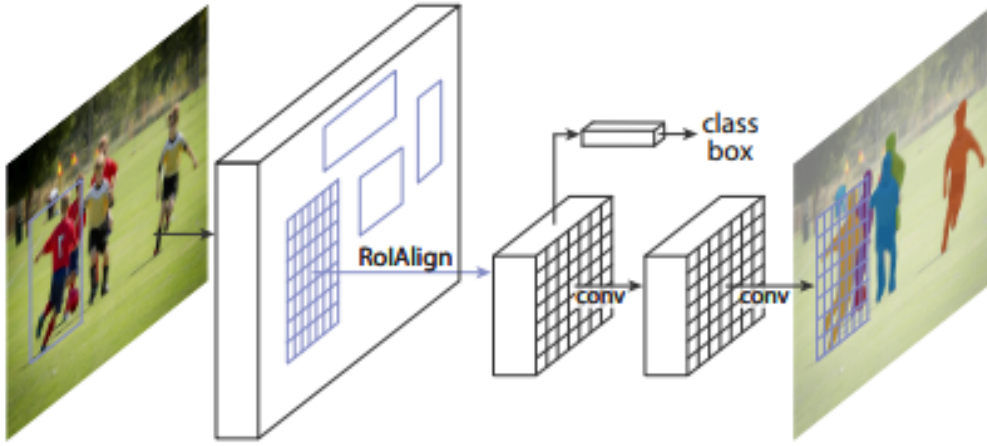


FIGURE 3.2: Mask R-CNN is a flexible and general framework for object detection and instance segmentation. It belongs to the family of region-proposal networks and an extension of Faster R-CNN. Objects in an image are detected, while simultaneously predicting high-quality segmentation mask at pixel-level, for each detected object. This instance segmentation achieves better localisation (He et al., 2017). Our two-stage mitosis detection framework, uses Mask R-CNN in the first stage to predict masks for the datasets.

3.1.1.2 Deep Feature Approaches

The deep neural network (DNN) approaches for mitosis detection and recognition can be classified as pixel-based, region-based, and hybrid approaches.

Pixel-level classifiers have been used for automatically analyzing histopathological images for cancer diagnosis (Cireşan et al., 2013). Patches obtained using a sliding window technique are classified as either mitotic or non-mitotic, based on the pixel classification of the centroid of a patch. For a given image, the pixels are assigned to one of the two classes, namely, mitotic, and non-mitotic. Then, a window is defined for each pixel, and an image patch corresponding to this window is used to train a DNN for pixel-level classification. The DNN has convolution layers, max-pooling layers, a fully connected layer, and a softmax layer at the end, for classification. Pixel-level approaches are computationally very expensive.

The studies (Saha et al., 2018) and (Wang et al., 2014a), proposes a combination of hand-crafted features with learned deep features for mitosis detection. In their study, (Saha et al., 2018) converted the original RGB image to a blue-ratio image and trained a DNN using patches around the detected objects, and combined it with hand-crafted features, for

detecting mitosis cells. Similarly, the study (Wang et al., 2014a), uses a pre-processed image, and combines results of several classifiers, trained separately on hand-crafted features and learned deep features. Both these approaches rely on domain-specific pre-processing steps for region-based feature extraction, to achieve good mitosis detection results.

In other studies (Li et al., 2018c) and (Chen et al., 2016), a two-stage process for mitosis detection has been proposed. In the first stage, coarse features were obtained and then refined in the second stage. In the first stage, the study (Li et al., 2018c), used random forests to obtain coarse features while the study (Chen et al., 2016), used a DNN. Both these methods, obtained fine features using a DNN in the second stage. These approaches are more robust to differentiating the actual mitotic cells from their mimics, but still perform poorly in localizing the mitosis cells, thus affecting the overall detection accuracy.

Approaches based on region proposals have also been proposed for mitosis detection. Region-based approaches as shown in Figure 3.3, are more efficient, as they process groups of pixels, which share similar attributes, such as color and texture. Region-based approaches rely on good segmentation for mitotic detection. The study (Li et al., 2018a), uses Faster R-CNN (Ren et al., 2015) for predicting bounding boxes around the mitosis cells. Faster R-CNN uses region proposals, where several anchor boxes at various scales are generated and mapped to regions for object detection. Fixed length features for each of the proposed regions are then used for both classification and localization. However, as the method by (Li et al., 2018a) does not predict masks for the cells, it performs poorly on mitosis cell localization. It is established that predicting accurate masks simultaneously with bounding boxes, can improve the accuracy of localization (He et al., 2017). Hence, with improved cell localization, cell-level features can be analyzed to identify mitosis cells with greater accuracy.

3.1.2 Contributions

In this chapter, we address the challenge of imprecise localization and expensive annotation required for instance segmentation, as outlined in Section 1.5.1. This chapter studies the consequences of insufficient localization on tasks such as object detection and classification. In this study, we have designed a cascaded detection method, by first generating

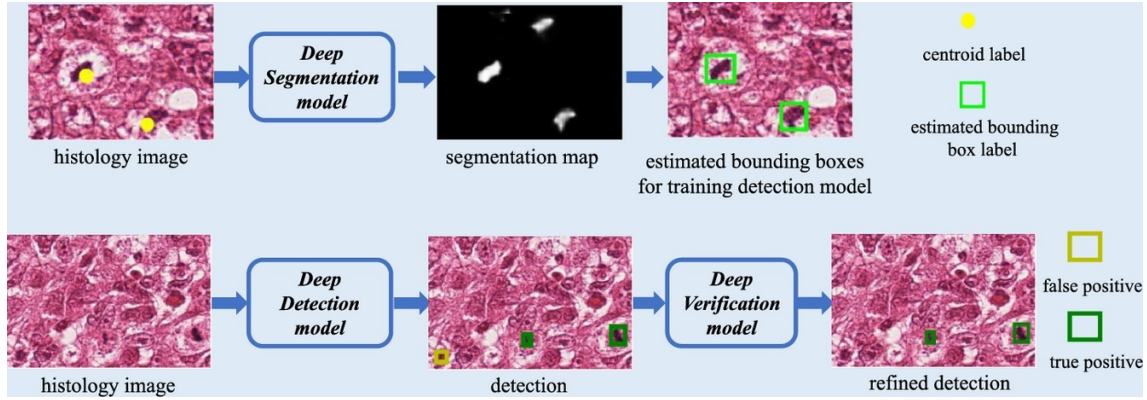


FIGURE 3.3: DeepMitosis uses a 2-step mitosis detection approach (Li et al., 2018a). A deep segmentation model based on FCN is used to generate candidate bounding boxes from weakly annotated datasets. Then a Faster R-CNN network is used in the deep detection model to predict bounding boxes, followed by deep verification step, which performs mitosis classification.

candidate segmentation masks using Mask R-CNN (He et al., 2017), and further using these masks to obtain learned deep features and hand-crafted features, to identify true mitotic cells. Mask R-CNN has a multi-task loss function, which allows the model to simultaneously learn object classification, object bounding box and an object mask. Further, by using transfer learning from the trained Mask R-CNN models, our method is capable of obtaining automatic segmentation for weakly annotated datasets such as ICPR Mitosis 2014¹ which lack ground truth masks. The multi-task learning improves the classification accuracy of the model. Further, with such mitosis mask generation, cell candidates are detected based on the confidence scores with a high recall rate for further processing. Our evaluation shows that our method outperforms the state-of-the-art approaches on both ICPR 2012 and 2014 mitosis datasets.

3.2 Method

Our mitosis detection framework has two stages. In the first stage, the mask generation stage, a trained mask generator network is used to produce masks for candidate mitosis cells along with confidence scores. In the second stage, the mitosis detection stage, the outputs of the first stage are refined to eliminate false positives using both learned deep

¹Mitos and Atypia 14 <https://mitos-atypia-14.grand-challenge.org/Dataset/>

features, and hand-crafted features extracted based on the masks generated in the first stage.

3.2.1 Mask Generation Stage

Recently, Mask R-CNN (He et al., 2017) has become one of the widely used state-of-the-art instance segmentation frameworks. Mask R-CNN uses a region proposal network (Ren et al., 2015) in the first stage to identify candidate bounding boxes and in the second stage predicts regressed bounding boxes, classification for the bounding boxes, and masks for the objects. The instance segmentation by Mask R-CNN is shown in Figure 3.2. Mask R-CNN uses RoI-Align layer to extract fixed length features, which are exactly aligned with the input. The extracted features are used for bounding box regression, classification, and prediction of masks within the bounding boxes using a separate branch of the network. Mask R-CNN has the advantage of predicting masks along with classification, and can be adapted for mitosis detection, where the morphological features of the cells have distinctive characteristics for the mitotic cells.

Specifically, we adopted the Mask R-CNN network to detect mitosis cell candidates with a high recall rate (possibly many false positives), as shown in Figure 3.4. We used ResNet-50 (He et al., 2016) as a backbone for feature generation and binary classification between foreground cells and background. The network was trained using high power fields (HPFs) from 40x magnification whole slide images with their masks. In this study, each HPF image was split into four non-overlapping images, which are then used to train the network. The anchor box sizes for Mask R-CNN were empirically determined by clustering the height and width of the ground truth bounding boxes in the training set. K -means clustering was used and the anchor box scales of 5, 15, 30, and 45 were used. The model predicted masks for mitosis cells along with confidence scores and bounding boxes.

To predict mitosis cells on the test set, a minimum confidence threshold of 0.8 and 40 detections per image were used, which were determined empirically, to obtain a 95% recall rate. This ensured that most mitosis cells had been detected along with a large number of false positives. These false positives were mitotic cell mimics classified as

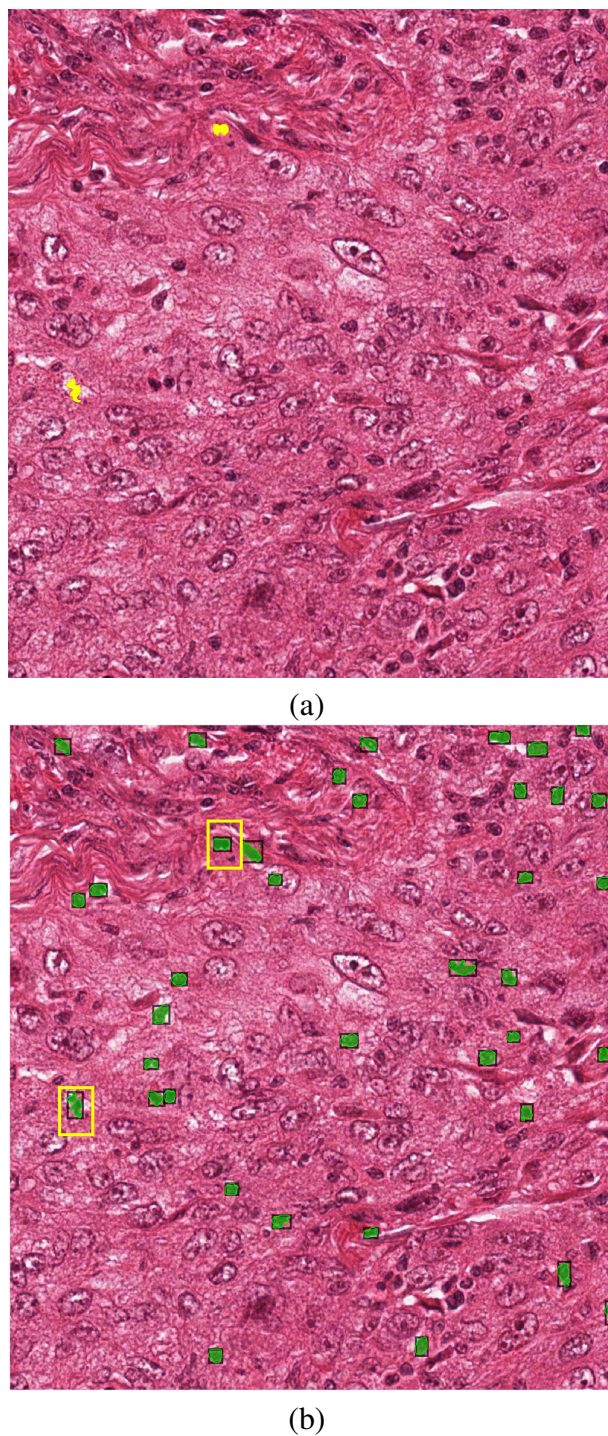


FIGURE 3.4: Illustration of mask generation using Mask R-CNN: (a) The image shows the High Power Field image with ground truth annotation in yellow; (b) The image shows masked mitosis cells in green, predicted by Mask R-CNN. The manually marked bounding boxes in yellow show the mitosis cells matching the ground truth mitosis cells. Other non-marked cells are hard negatives mitosis cells which are false positives.

mitotic with a high confidence score. The detection outputs of the first stage were then processed further to eliminate false positives in the second stage.

3.2.2 Mitosis Detection Stage

Inputs to the second stage were the masks for the mitosis cells, bounding boxes and prediction confidence scores obtained from the first stage. The predicted masks were used to generate local patches of 72×72 pixels around the detections. Figure 3.6 illustrates the mitosis detection stage.

In this stage, learned deep features and hand-crafted features were extracted from the local patches. To derive the learned deep features, the Xception network (Chollet, 2017) pre-trained on ImageNet (Deng et al., 2009) without the last two fully connected layers was used. The Xception network architecture is shown in Figure 3.5. Hence, this deep architecture helps to capture texture, morphological patterns. The small parameter size of the Xception network with deep separable convolution networks, helps to reduce over-fitting so that the trained model can be used on other datasets to obtain good mitosis detection accuracy. Hence, the network generalizes better. To further improve its discriminative power, the learned deep features from the last convolution layer were combined with hand-crafted features to incorporate domain knowledge, for better classification.

Our implementation used hand-crafted features which included morphological, intensity, texture features such as area, perimeter, orientation, aspect ratio, angle magnitude and gradient histogram, RGB colour histogram, Haralick, local binary patterns, Zernike moments, eccentricity, threshold adjacency statistics, and entropy. In addition, we also used the boundary length of the mask and the radius of the minimum enclosing circle of the mask. The combination of learned deep features and hand-crafted features mimics the strategy used by pathologists for identifying mitosis cells. This strategy involves identifying mitosis cells based on learning from other exemplars combined with identifying certain auxiliary features of mitosis cells such as hyper-chromaticity, hairy extensions and shapes (Veta et al., 2014). This concatenated feature vector of hand-crafted features of dimension 278, and learned deep features, was used to train the two fully connected (FC) layers, with the last layer having softmax activation for classification (see Figure 3.6). All

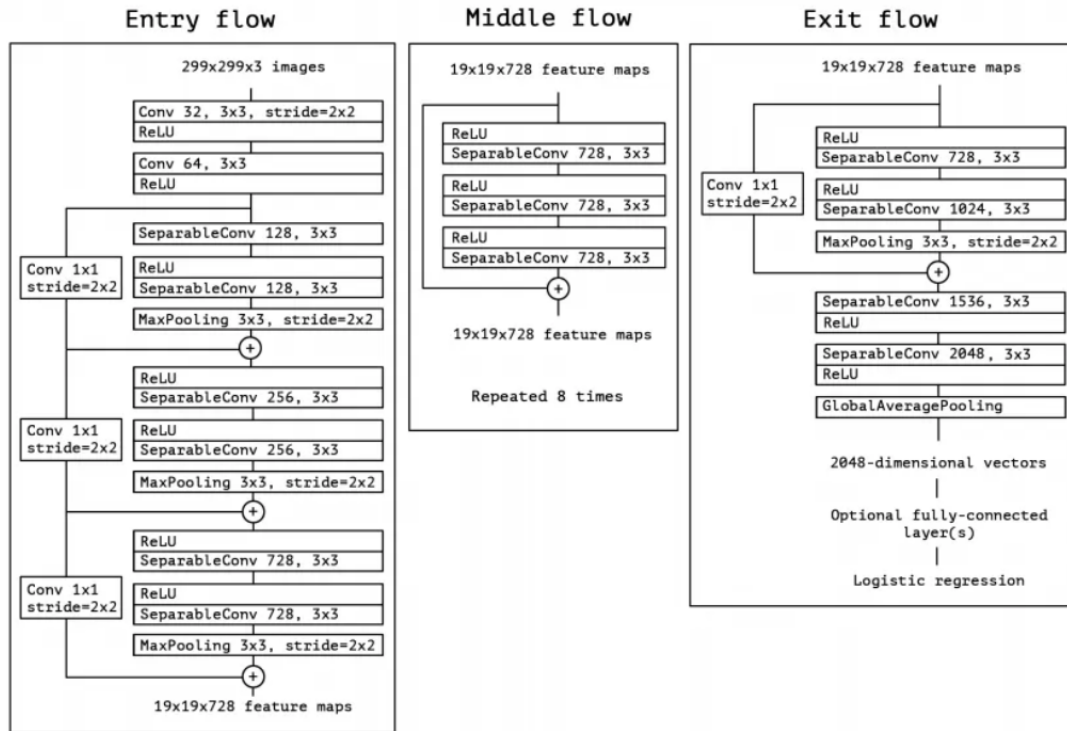


FIGURE 3.5: Xception network (Chollet, 2017) which is used for the mitosis detection stage. Xception network has depth-wise separable convolution layers. The network has three distinct parts, entry flow, middle flow and exit flow. The first two are repeated eight times. Hence, the training data goes through eight sets of entry and middle flow parts before it is passed into the single exit flow part. For the mitosis detection stage, we use a modified version of this network where the last two fully connected layers and the classification layer are removed.

other layers used ReLU as the activation function instead of tanh, to avoid the vanishing gradient problem. Categorical cross-entropy loss was used during training.

As only 40 predictions per image were used from the mask generation stage, training was done using only hard negatives. During training, the data was shuffled in the batches such that in each batch 90% was positive samples and 10% negative samples; so that the network would learn more about the positive features and then differentiate that from hard-negative samples. Augmentation methods such as rotation with a range from 0 to 40 degrees, horizontal and vertical flipping were used to compensate for the small dataset, to prevent over-fitting. The positives of predicted output of the mitosis detection stage were re-classified as negatives based on the confidence score obtained from the mask generation stage.

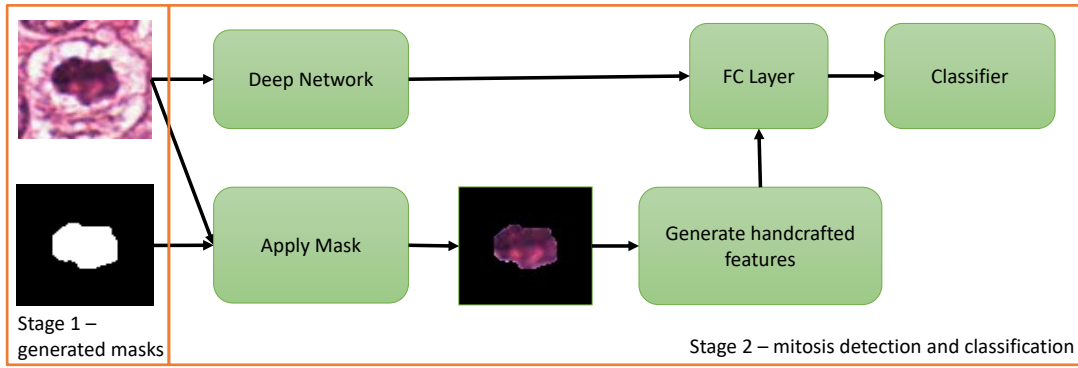


FIGURE 3.6: Mitosis detection stage where candidate detections from stage 1 are classified as mitotic or non-mitotic. Hand-crafted features extracted from the masks generated from the Mask R-CNN network are combined with deep features to classify the candidate cells. The input images are RGB image and the binary mask image. The learned deep features are extracted using the Xception network or the VGG-16 network. The combination of learned deep features and hand-crafted features are used to train a network with a fully connected layer followed by a mitosis classification layer.

3.3 Experiments and Results

3.3.1 Datasets

Experimental evaluation was done on the ICPR 2012 (Roux et al., 2013) and ICPR 2014² mitosis datasets. Figure 3.7 provides examples of image patches with mitosis cells, extracted from the two datasets. Both these datasets contain breast cancer biopsy slides, where the slides are stained with standard hematoxylin and eosin (H&E) dyes, and scanned by one of two slide scanners, namely, Aperio Scanscope XT and Hamamatsu Nanozoomer 2.0-HT.

ICPR 2012 dataset: The dataset has 50 High Power Field (HPF) images, sampled from breast cancer histopathology images at $\times 40$ magnification, out of which 35 HPF images were used for training and 15 for testing. The images have a resolution of 2084×2084 pixels for each field, contain 101 mitosis cells in the test set and 226 in the training set. For this study, we considered the images from the Aperio XT scanner only. ICPR 2012 dataset has mask annotation for all the mitosis cells.

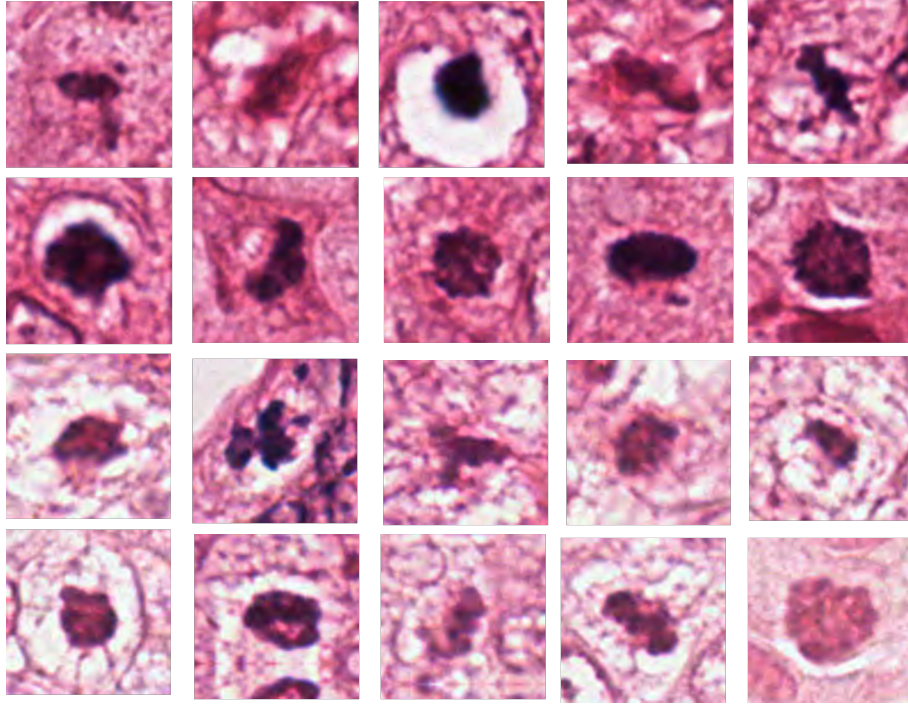
²Mitos and Atypia 14 <https://mitos-atypia-14.grand-challenge.org/Dataset/>

ICPR 2014 dataset: The dataset has 1200 High Power Field (HPF) images, sampled from breast cancer histopathology images at $\times 40$ magnification. The corresponding test dataset has 496 images, but were not considered for this study, since they lacked annotations. All the images have a resolution of 1539×1376 pixels for each field and contain 610 mitosis cells. The training dataset of 1200 image was split into training and validation sets following the same split protocol as mentioned in (Li et al., 2018a), so that we could make a fair comparison. For this study, we considered the images from the Aperio XT scanner only.

ICPR 2014 dataset is annotated with only centroids of mitosis cells, hence it is weakly annotated. Hence, for segmentation task the dataset can be considered as being weakly annotated. For the purpose of this study, we also chose to evaluate on the ICPR 2014 dataset to demonstrate that our model trained on a small dataset like ICPR 2012, could be used to generate reliable masks for weakly annotated datasets such as ICPR 2014. In addition, because the ICPR 2014 dataset does not provide ground truth for the testing dataset, we used a subset of the training dataset in our evaluation and excluded that subset for training the model as explained earlier. The training set for ICPR 2014 dataset for mitosis detection stage had 570 mitotic cells and the validation set had 139 mitotic cells. The training set contained fewer mitosis cells than actual ground truth 610 mitotic cells, since all the mitotic cells were not predicted by the Mask R-CNN network.

3.3.2 Implementation and Results

The Mask R-CNN network was trained for 30k epochs, with stochastic gradient descent as the optimizer with a learning rate of 0.001. The network was pre-trained on ImageNet. Vertical and horizontal flip and 0 to 40 degrees rotation range augmentations were used. The training time for Mask R-CNN was 6s per epoch. For the mitosis detection stage, we trained an Xception network for 1500 epochs with Adam optimization and an initial learning rate of 0.001. The training time was 3s per epoch. The experiment was conducted on Xception network as well as VGG-16 network (Simonyan and Zisserman, 2014). An empirically determined threshold of 0.9 on the confidence score was used on the predicted output to further reduce the false positives. All experiments were conducted on a NVIDIA



(a)

FIGURE 3.7: Sample image patches from the Mitosis 2012 and 2014 dataset showing mitosis cells. The first two rows correspond to Mitosis 2012 dataset and the last two rows are from the Mitosis 2014 dataset. Structural similarity can be noticed in the cells from the two datasets, but the overall appearance differs due to non-standard staining and tissue preparation process.

GeForce GTX Titan dual-GPU machine, and the code was implemented using Keras and Tensor Flow.

Evaluation was done as per the criteria in the ICPR challenges. Precision, recall and F1-score were used as evaluation metrics³. Mitosis detections were considered correct if a detection lay within 20 pixels and 32 pixels from the ground truth centroid pixel for ICPR 2012 and ICPR 2014 datasets respectively, which was according to the ICPR contest rules.

Table 3.1 provides the results obtained on the ICPR 2012 dataset. Our method using the Xception network has outperformed the best state-of-the-art DeepMitosis method (Li et al., 2018a) in both F1-score and precision. Furthermore, our method using VGG-16 network performs similar to (Li et al., 2018a), with same recall and a similar F1-score. DeepMitosis (Li et al., 2018a) uses Faster R-CNN (Ren et al., 2015) which uses region

³Challenge Evaluation Metrics <https://mitos-atypia-14.grand-challenge.org/Evaluation-Metrics/>

TABLE 3.1: Results of our two-stage mitosis detection framework in two different settings and comparison with state-of-the-art methods on the ICPR 2012 dataset. Our method, with the Xception network in the second stage, achieves the best precision and F1 score, and with the VGG-16 network, achieves the equal-best recall score.

Method	Precision	Recall	F1-Score
DeepMitosis (Li et al., 2018a)	0.85	0.81	0.83
HC+CNN (Wang et al., 2014a)	0.84	0.65	0.73
Pixel-Based (Cireřan et al., 2013)	0.88	0.70	0.78
RCNN (Li et al., 2018c)	0.78	0.79	0.78
Ours (VGG-16)	0.87	0.81	0.84
Ours (Xception)	0.93	0.80	0.87

proposal network for localizing mitotic cells. The detections are further refined by using a ResNet based classification model to remove false positives and the segmentation network is used to obtain bounding boxes for weakly annotated datasets such as ICPR 2014 datasets where only the centroid of the cell is annotated. Our method also outperforms (Cireřan et al., 2013), (Wang et al., 2014a), and (Li et al., 2018c). The approach (Cireřan et al., 2013), uses a pixel-based classification. The study (Wang et al., 2014a), uses a cascade approach of combining hand-crafted and learned deep features for mitosis detection and a very light CNN network for classification. RCNN (Li et al., 2018c) uses hand-crafted features to detect the candidate mitosis cell regions which are refined using a lightweight region-based CNN and spiral scan for mitosis pixel classification.

Our method outperforms other methods because it captures the spatial information using a deep neural network and also captures localized mitosis cell features using accurate masks and tight bounding boxes. The accurate masks also help to compute robust hand-crafted features than those methods without masks (Wang et al., 2014a). Our method provides an advantage over pixel-based approaches (Cireřan et al., 2013) which use a fixed bounding box around the mitotic pixel. Methods such as RCNN (Li et al., 2018c) use both learned deep features and hand-crafted features, but their features are noisy as they do not use predicted masks to localize mitosis cells. One of the main differences between our

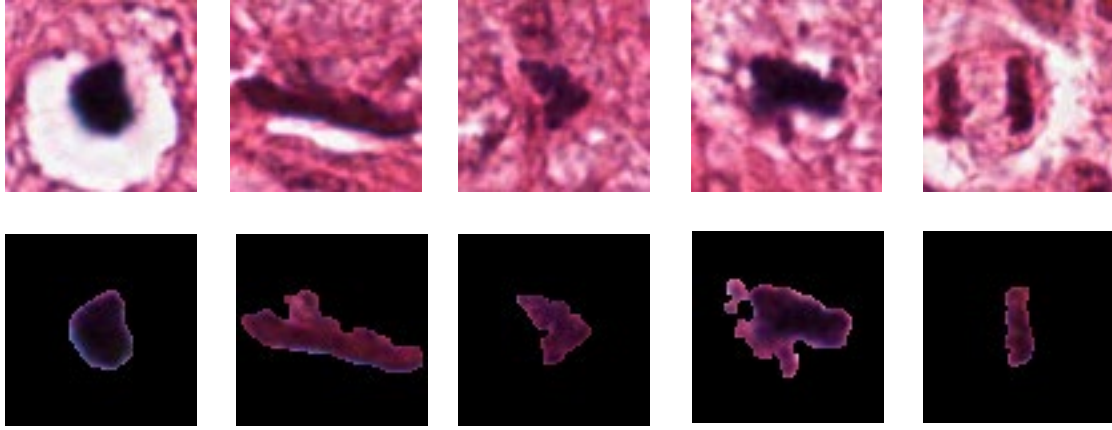


FIGURE 3.8: Masks generated for sample image patches from the weakly annotated ICPR 2014 Mitosis dataset by our Mask R-CNN model trained on ICPR 2012 dataset. The first row contains the image patches, and the second row has the corresponding masks applied to the image patches to isolate the cells. Note that the masks generated are accurate in all cases except the fifth image from the left, where the cell is in the telophase stage.

approach and DeepMitosis is that we do not have a separate segmentation and detection network. They are combined into a single mask generation and detection network based on Mask R-CNN. Furthermore, DeepMitosis also does not use hand-crafted features with masks to get localized cell features.

With the promising results on the ICPR 2012 dataset, we were interested to find out if the model trained on the ICPR 2012 dataset for generating masks could also be used to generate masks for other datasets. To evaluate this, we generated the masks for the ICPR 2014 dataset using the Mask R-CNN model trained on the ICPR 2012 dataset, and then performed mitosis detection on the validation dataset using both VGG-16 and Xception networks. Table 3.2 provides the performance comparison of the various methods on the ICPR 2014 validation dataset where the validation dataset was similar to (Li et al., 2018a) in terms of positive and negative samples chosen. Figure 3.8 shows the masks generated, with the top row depicting the cells of Mitosis ICPR 2014 dataset in their original stained background. They are weakly annotated. The second row shows the masks generated from the model. The results show that the VGG-16 model achieved a better F1-score and recall than DeepMitosis (Li et al., 2018a) and RCNN (Li et al., 2018c), while RCNN has a better precision when compared to our method due to the scanning operation they use to find mitotic cells. This demonstrates that the trained Mask R-CNN model is highly

TABLE 3.2: Results of our two-stage mitosis detection framework in two different experimental settings and comparison with state-of-the-art methods on the ICPR 2014 dataset where evaluation was performed on the ICPR 2014 validation set. Our method based on VGG-16 outperforms state-of-the-art DeepMitosis and RCNN in both recall and F1 scores. Our implementation, based on Xception, achieved the second-best recall score.

Method	Precision	Recall	F1-Score
DeepMitosis (Li et al., 2018a)	-	-	0.57
RCNN (Li et al., 2018c)	0.65	0.66	0.66
Ours (Xception)	0.62	0.67	0.64
Ours (VGG-16)	0.58	0.82	0.68

transferable despite the small size of the ICPR 2012 dataset. This further suggests that in future, if we could obtain a Mask R-CNN model trained on a large dataset we could possibly use it to perform reliable mitosis detections on weakly annotated datasets. In our method we think that the inaccuracies of the mask prediction on the ICPR 2014 dataset might have contributed to the relatively low accuracy of mitosis detection using the Xception network. This needs to be further investigated by generating accurate masks for the ICPR 2014 dataset. We also plan to combine the two stages to provide an end-to-end solution for mitosis detection.

3.3.3 Ablation Experiments

3.3.3.1 Simpler Network and Deep Features

For the mitosis detection stage, we first experimented with simpler networks. The entire setup of the experiment was exactly as described in the Section 3.2.2. The only difference is that instead of VGG and Xception network, simple convolution network was used. The network has the configuration of five alternating convolution layers of filter sizes 2, 4, 3, 2, and 2, and max pooling layers of filter size 2, and two fully connected layers, one with 1000 units and the another with 2 units for softmax classification loss. The network was trained for images of input size 72×72 . The result of the experiment is highlighted in Table 3.4. We noticed that the result was not promising and could not capture the

TABLE 3.3: Results of ablation experiments of our two-stage mitosis detection framework comparison with state-of-the-art methods on the ICPR 2012 dataset. In particular, when we used learned deep features only to train our mitosis detection stage, our method matched DeepMitosis in F1 score and showed marginal improvement in precision score. When using deep learned features, the performance was only marginally lower when compared to our best result which used both learned deep features and hand-crafted features (first two rows).

Method	Precision	Recall	F1-Score
Ours (VGG-16)	0.87	0.81	0.84
Ours (Xception)	0.93	0.80	0.87
DeepMitosis (Li et al., 2018a)	0.85	0.81	0.83
Ours (Xception) Deep features only	0.86	0.80	0.83

morphological and textural features of the cells. We then extended the experiment to deeper networks like Xception and VGG and found that it was better suited for mitosis detection. The Xception network has depth-wise separable convolution network. The reduced parameters due to this architecture, makes it generalize better than the simple convolution network.

Table 3.3 shows the results of our implementation, using just learned deep features. The result of using only learned deep features was similar to DeepMitosis (Li et al., 2018a), where only deep features were used. However, this result was lower than our best result of using both hand-crafted and deep features.

3.3.3.2 Faster R-CNN for Detections

In methods such as DeepMitosis, Faster R-CNN was used for detections. We wanted to compare the results when Faster R-CNN was used instead of Mask R-CNN. Faster R-CNN architecture predicts the bounding boxes with region proposals and anchor boxes of various scales. To conduct our experiment fairly, we used the same architecture set-up and replaced the Mask R-CNN with Faster R-CNN and obtained the bounding boxes. The features were extracted from the bounding boxes. Our hypothesis was confirmed from the results, that Faster R-CNN’s performance was lower than that of Mask R-CNN, and the

TABLE 3.4: Results of the ablation experiments of our two-stage mitosis detection framework, and comparison with state-of-the-art methods, on the ICPR 2012 dataset. In particular, we compared the results of using a simple network with 5 convolutional layers, with VGG-16 and Xception network. Across all metrics, the results were lower.

Method	Precision	Recall	F1-Score
Ours (VGG-16)	0.87	0.81	0.84
Ours (Xception)	0.93	0.80	0.87
Ours (5 Conv layers)	0.64	0.75	0.69

F1 score was 0.81 when compared to 0.87 of that of Mask R-CNN. This experiment further demonstrates that Mask R-CNN provides better localisation, which leads to better features for classification.

3.4 Summary

In this chapter, we have proposed a method of mitosis detection by generating masks automatically for the mitosis cells and its close mimics, and then classifying the cells using learned deep features and hand-crafted features. Predicting masks simultaneously with bounding boxes helps improve localization of mitosis candidates. Furthermore, the predicted masks facilitate the localized extraction of morphological and texture features of the mitosis candidates, which augment the learned deep features in the mitosis detection stage. The combination of learned deep features and the hand-crafted shape and texture features mimics the strategy used by pathologists for identifying mitosis cells. Evaluation on the ICPR 2012 dataset and ICPR 2014 training datasets shows that our method outperformed the state-of-the-art approaches. Our experiments also show that our model trained on the ICPR 2012 dataset can be effectively applied to the ICPR 2014 dataset, which is weakly annotated, and achieves good detection results. Thus, we have demonstrated that our method provides promising transferability across datasets.

Part 2

Advanced Learning-based Weakly Supervised Methods

Dual-Stage Domain Adaptive Mitosis Detection for Histopathology Images

The deep learning models trained on a dataset from one domain generally perform poorly on datasets from another domain as the learned feature representations have different distributions. Hence, they are faced with the domain-shift problem and suffer from poor generalization caused by poor domain alignment. One way to mitigate this problem is to train the model containing representative data from all domains. However, this approach is expensive and impractical. Many domain-adaptive approaches have been discussed in Section 1.2.2 to address the domain-shift issue, where models are trained such that, the source domain data is supervised, and the target domain is unsupervised. However, none of these methods provide domain alignment of features for small objects with different backgrounds and morphological characteristics.

Histopathology images for mitosis detection vary in appearance due to the non-standard method of preparing the tissues as well as differences in scanner hardware. This makes automatic machine learning based mitosis detection very challenging, because of the domain shift between the training and testing datasets. There have been number of methods proposed to handle this domain shift, such as stain normalization, augmentation, and supervised domain adaptation. But these methods are not robust, rely on large, annotated datasets or pre-processed datasets. In this chapter, we propose a method of addressing this domain shift problem, by using a two-stage domain adaptive neural network. In the first stage, we use domain adaptive Mask R-CNN to generate masks for mitotic regions. The generated masks are then used by a second domain adaptive convolutional neural network to perform finer mitosis detection. Our method achieved state-of-the-art performance on both ICPR 2012 and 2014 datasets. We demonstrate that, using a domain agnostic approach, achieves better generalization and mitosis cell localization, for

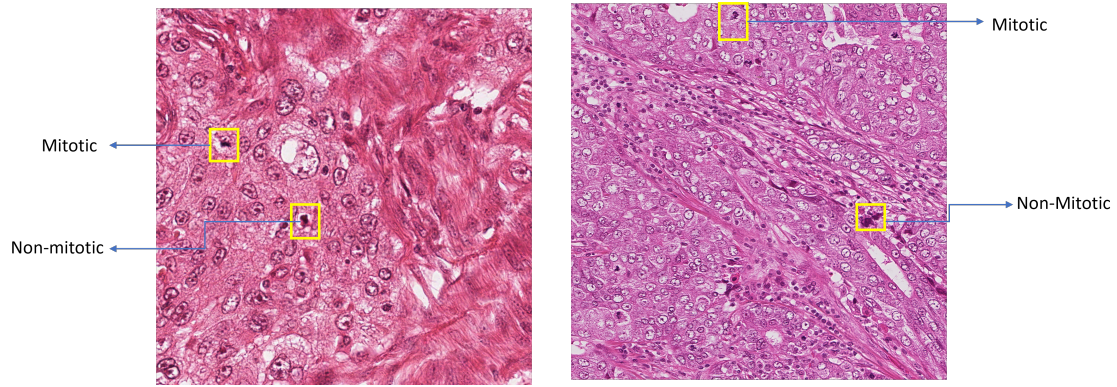


FIGURE 4.1: Effect of non-standard tissue preparation and staining. (Left), a whole slide image from ICPR 2012 dataset and (right) a whole slide image from ICPR 2014 dataset. In both instances, examples of mitosis and non-mitosis cells are shown using yellow bounding boxes.

the trained models. The content of this chapter is related to the publication (Dodballapur et al., 2020b).

4.1 Introduction

With the advent of fast slide scanners, pathologists now look for mitosis cells for cancer diagnosis. The tissues are stained with Hematoxylin and Eosin (H&E) stains to differentiate between the background cytoplasm and the nucleus. The high degree of variance in the appearance of the histopathology images is due to non-standard tissue preparation, and differences in scanning systems, both of which present a key challenge to be overcome for automatic mitosis detection, as shown in Figure 4.1. Deep neural methods have been used in literature in computer vision for many applications such as object detection, classification and semantic segmentation. A known limitation of deep neural networks (DNNs) is that a network trained with images with a particular tissue preparation, performs poorly on images with a different tissue preparation. This phenomenon is known in the literature as the domain shift problem. It is caused by a shift in the distribution of deep learned features due to a shift in the distribution of the input images between training and testing time. A similar problem of domain shift exists in other fields, such as in pedestrian and vehicle detection and recognition. For example, objects often need to be detected and recognized when they are under different viewpoints, backgrounds, lighting conditions during testing time when compared to training time.

One of the ways that domain shift can be tackled is to use training data with the same distribution as test data. This, however, requires more time and effort, as each time there is a change in test data, we have to collect and annotate the appropriate training data. Hence, there is a need to explore models that can easily handle these visual changes between the training and test dataset. The following section summarizes the related work in this field.

4.1.1 Related Work

In the literature, several approaches for domain adaptation have been proposed for handling the domain shift problem in the context of DNNs (Cariucci et al., 2017; Li et al., 2017a; Long et al., 2015b; Xu et al., 2014), particularly for image classification and object detection. The study (Xu et al., 2014), uses adaptive structural SVM, with very few samples of data from the target domain during training, to make the model domain adaptable while (Long et al., 2015b) proposes using multiple kernels to produce unbiased deep features. The study (Cariucci et al., 2017), proposes to align the representations of source and target domain using a reference domain while (Li et al., 2017a) proposes to generate domain shift features from deep neural networks. The study (Ganin and Lempitsky, 2014), introduces a gradient reversal layer in a deep network to produce deep features which are domain invariant. In particular, for object detection (Chen et al., 2018c), no annotations are required for the target domain data. In this method (Chen et al., 2018c), supervised training is done on the source domain data, but no supervision is available for the target domain data. A limitation of (Chen et al., 2018c) is that it does not predict masks for detected objects, which are shown to be beneficial for mitosis detection (Saha et al., 2018; Dodballapur et al., 2019).

In the context of mitosis detection, some of the methods for mitosis detection particularly do not handle domain shift (Li et al., 2018a; Wang et al., 2014a; Dodballapur et al., 2019; Li et al., 2018c; Tashk et al., 2013; Chen et al., 2016). An example of such a method is a region-based approach for mitotic detection (Li et al., 2018a). Another study (Dodballapur et al., 2019), proposes a mask-driven localization approach, which improves the mitosis detection accuracy, while (Li et al., 2018c) uses light weight region-based CNN and a spiral scan for mitosis pixel classification.

To address the problem of domain shift, stain normalization (Macenko et al., 2009; Ruifrok et al., 2001) can be performed on the images as a pre-processing step before training a neural network (Hamidinekoo and Zwigelaar, 2017). While this method improves on the performance of the deep network despite variations in input images due to stain colour, it does not address variations due to processing of the tissue to preserve its underlying components and cells. To overcome the limitations of (Hamidinekoo and Zwigelaar, 2017), a supervised domain adaptation method (Lafarge et al., 2017) was proposed to use annotations for both the source and target domains to train a DNN to learn domain invariant features. Such a method requires fully annotated data for the target domain which can be expensive to obtain. Augmentation is another popular technique used to handle different stains and conditions. Examples include a method of stain augmentation, which modifies the H and E colour channels instead of R,G,B (Tellez et al., 2018). However, augmentation is not robust to all the different variations in tissue processing.

4.1.2 Contributions

In this chapter, building on the supervised localization methods of first part of thesis, we delve into tackling the challenge of domain-shift which has been outlined in Section 1.5.3, which is caused by different training and testing domains. While many methods in Section 1.5.3, have been promising in solving domain-shift problem, they have not been that successful in solving the domain alignment problem of small object feature representations, with varied morphological features and staining backgrounds.

The main contribution of this chapter is a dual-stage domain adaptive mitosis detection framework which has the advantage of learning domain invariant features at both image and instance levels. The motivation behind using a two-staged approach is that mitosis cells are very similar to other cells present in histopathology images. So, a cascade approach gives the flexibility of first getting a coarse detection, and then a second stage of finer detection, to eliminate the mitosis cell mimics. Further, the two-stage approach helps us to focus on different aspects of domain adaptation at each stage. In our method, supervised training is done on the source domain data, but no supervision is available for the target domain data. Both stages of our pipeline include a domain adaptation component based on the H -divergence theory, which learns a domain classifier in an adversarial

manner. In the first stage, we generate instance level localized masks using adversarial training, to detect candidate mitosis cells with a high recall. In the second stage, we further process the masks of mitosis cells as well as their mimics (hard negatives which closely resemble mitosis cells), to improve the precision of the mitosis detection by learning image level domain invariant features.

Our approach is novel because our two-staged convolutional neural network uses domain adaptation in both stages unlike both (Dodballapur et al., 2019) and (Chen et al., 2016). Domain adaptation is not considered in (Dodballapur et al., 2019) and (Chen et al., 2016) while the main focus of our chapter is domain adaptation in both stages, with each stage exploring a different kind of domain adaptation, namely image level (second stage) and instance level (first stage) domain adaptation. Hence, the domain adaptation introduced in both stages makes it sufficiently novel and unique.

Our domain adaptation method is supervised in the source domain and unsupervised in the target domain unlike (Lafarge et al., 2017) and uses a stage specific domain adaptation, which is not taught in the literature. The study (Lafarge et al., 2017) uses adversarial update on the mitosis classification loss, while our method applies it to the domain loss resulting in superior performance. Furthermore, (Lafarge et al., 2017) uses labels for both source and target domains while our method uses no labels for the target domain. Finally, unlike (Lafarge et al., 2017), we do not use colour augmentation.

Our approach differs from (Chen et al., 2018c) in that ours is dual-stage, where we introduce domain adaptation at both stages with a different focus of domain adaptation at each stage. In the first stage, domain adaptation is performed at the bounding box level, while in the second stage, it is performed at the image level. Further, our variation in domain adaptation strategy between the first and second stages addresses problems specific to detection and classification of mitosis cells with variations in morphological characteristics and staining.

We conducted experiments to evaluate the robustness and accuracy of masks as well as the overall mitosis detection accuracy using the ICPR 2012 (Roux et al., 2013) and ICPR

2014¹ datasets. Our evaluation showed the advantage of learning domain invariant features over non-domain invariant features for mitosis detection.

4.2 Method

Our method for mitosis detection is a pipeline with two stages namely the domain adaptive mask generation stage and the domain adaptive mitosis detection stage. For the first stage, inspired by the work of (Dodballapur et al., 2019), we modify the Mask R-CNN to predict masks and bounding boxes using domain invariant features. The classification output and masks from this stage are input into the detection stage, where classification of the detected regions into mitotic and non-mitotic cells is further refined by removing a large number of false positives, which correspond to mitosis cell mimics.

4.2.1 Domain Adaptive Component

The goal of our domain adaptive network is to minimize the H -divergence (Ben-David et al., 2010) between the source and target domains. This is achieved by making the learned feature distributions from the two dissimilar input domains as similar as possible, by using adversarial training. In both stages of our framework, we introduce a domain adaptive component. In essence, the domain adaptive component has a gradient reversal layer (GRL) (Ganin and Lempitsky, 2014) followed by two 1×1 convolution layers. The final softmax layer predicts the domain label probability for each feature element on the last convolution layer. The gradient reversal layer uses a negative training signal which aligns the distribution of the features of the source and target domains at the layers of the network before the gradient reversal layer. This alignment helps to handle the domain shift problems caused by changes in colour, lighting, staining and tissue preparation, which are common in histopathology images. Figure 4.2 and Figure 4.3 show the two stages of our pipeline where each DNN has a domain adaptive branch. The gradient reversal layer in our mitosis detection is placed between the feature extractor and the domain classifier component. While the domain classifier loss is minimized, the negative gradient in the back propagation helps train the feature extractor to maximize the domain classifier loss.

¹Mitos and Atypia 14 <https://mitos-atypia-14.grand-challenge.org/Dataset/>

4.2.2 Domain Adaptive Mask Generation

The first stage of our dual-stage mitosis detection pipeline is a mitosis mask prediction stage. In this stage, we use Mask R-CNN (He et al., 2017) with a ResNet-50 (He et al., 2016) backbone, pre-trained on ImageNet (Deng et al., 2009), as our baseline segmentation model. Mask R-CNN uses a region proposal network to identify candidate bounding boxes with regression and classification and masks for the objects. Since Mask R-CNN predicts masks, it is useful for mitosis detection to accurately localize the mitosis cells. By using the ROI-Align layer, Mask R-CNN provides the advantage of localizing mitosis cells (which are small in size) with greater accuracy when compared to Faster R-CNN, which uses the ROI-pooling layer, which introduces quantization errors in bounding box prediction. Furthermore, the authors of Mask R-CNN show that predicting masks improve bounding box accuracy as well.

We modified the Mask R-CNN network depicted in Figure 4.2, by adding a domain adaptive network branch, so that we could generate accurate and robust masks, by optimizing the loss function for mask generation and domain classification simultaneously. The domain adaptive component branches out of the last convolution layer of the ResNet-50 backbone, and the last softmax layer of the domain adaptive branch predicts the domain label probability for each feature element on the last convolution layer of ResNet-50 backbone. In this way, the adversarial domain classification helps the network to learn instance level domain invariant features, from which masks for mitosis regions are then predicted.

The inputs to the domain adaptive Mask R-CNN network are the high-power fields (HPFs) from whole slide images, masks representing mitosis cells (when available), and domain labels indicating the dataset the images belong to. In all our experiments, we used a domain label of 0 for the source domain, and domain label of 1 for the target domain. The network was trained using images from both the source and target domains. However, the ground truth masks are available only for the source domain, hence the training was supervised in the source domain and unsupervised in the target domain for the Mask R-CNN branch. The domain classification loss, which is a binary cross-entropy classification loss, was combined with the losses from the Mask R-CNN during training. This allowed the network to predict robust masks from domain invariant features.

We empirically determined the size of anchor boxes in Mask R-CNN by using cluster centers obtained from clustering the height and width of the bounding boxes in the training set. Specifically, using K -means clustering, we determined four dominant clusters which corresponded to anchor box sizes of 5, 15, 30, and 45. We used data augmentation methods such as rotation using range from 0 to 40 degrees, horizontal and vertical flipping, to prevent model over-fitting. The domain adaptive Mask R-CNN network was trained with stochastic gradient descent with a learning rate of 0.001. The trained Mask R-CNN model was then used to predict segmentation masks for candidate mitosis cells, along with their confidence scores and bounding boxes.

4.2.3 Domain Adaptive Mitosis Detection

The previous stage produces mitosis detection outputs with a high recall and a low precision. The second stage is thus designed to improve precision by reducing the number of false positives. Specifically, we propose a domain adaptive mitosis detection network with a VGG-19 network backbone pre-trained on ImageNet. To this network, we introduce similarly the domain adaptive component in addition to the mitosis classification branch, to generalize the model for classification as shown in Figure 4.3.

During training, the inputs to this stage are the predicted masks of mitosis cells, and hard negatives obtained from the previous domain adaptive mask generation stage. Hard negatives are mitosis cell mimics which are incorrectly classified as mitosis cells with a high confidence. In our implementation, we empirically determined this high confidence threshold to be 0.9. Local patches of size 72×72 around the centroid of the cell detections were extracted for training. For the source domain, we derive the label, true mitosis or not, for the predicted masks; but for the target domain we do not use any labelling for the mitosis classification branch. Data augmentation similar to the previous stage was used during training. A binary cross-entropy loss function was used to optimize the network. This network model was trained with Adam optimization and an initial learning rate of 0.001. During testing, HPF images were first input into the Mask R-CNN stage to obtain candidate mitotic regions. Local patches of size 72×72 around the centroid of the mitosis cell detections was then used to train the second refinement stage for final mitosis classification.

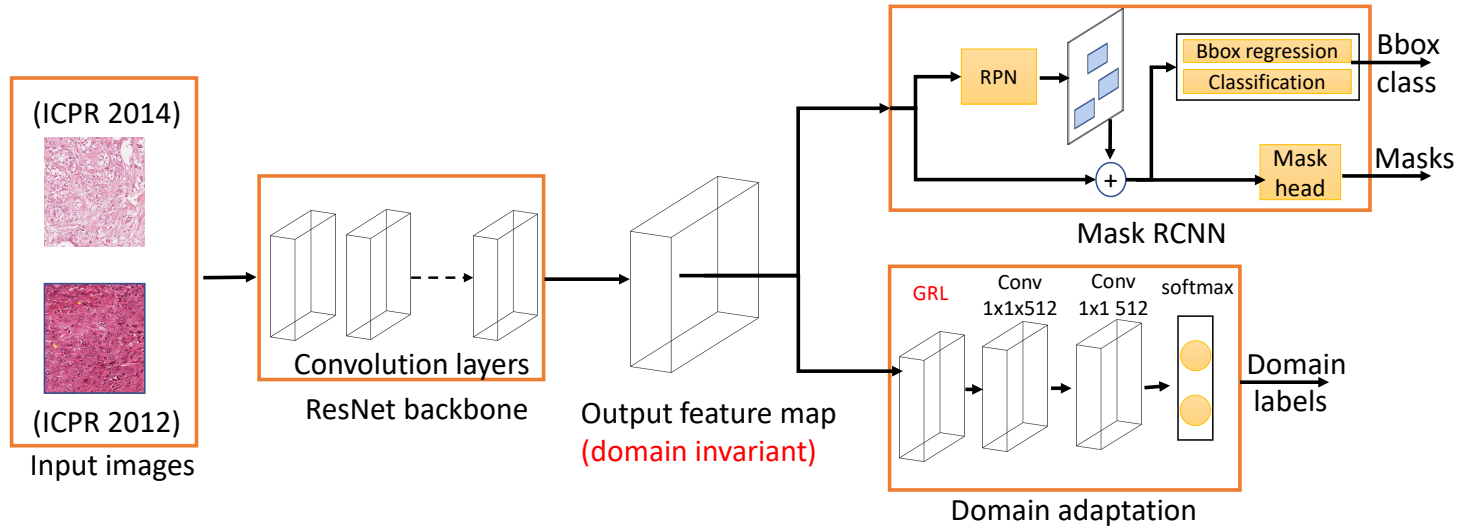


FIGURE 4.2: Domain Adaptive Mask Generation: the last layer of the ResNet-50 backbone connects both the Mask R-CNN head as well as the domain adaptation branch. A gradient reversal layer has been introduced to the Mask R-CNN network to enable the network to learn domain invariant features. The domain adaptive Mask R-CNN is used in the first stage of our dual-stage domain adaptive mitosis detection framework.

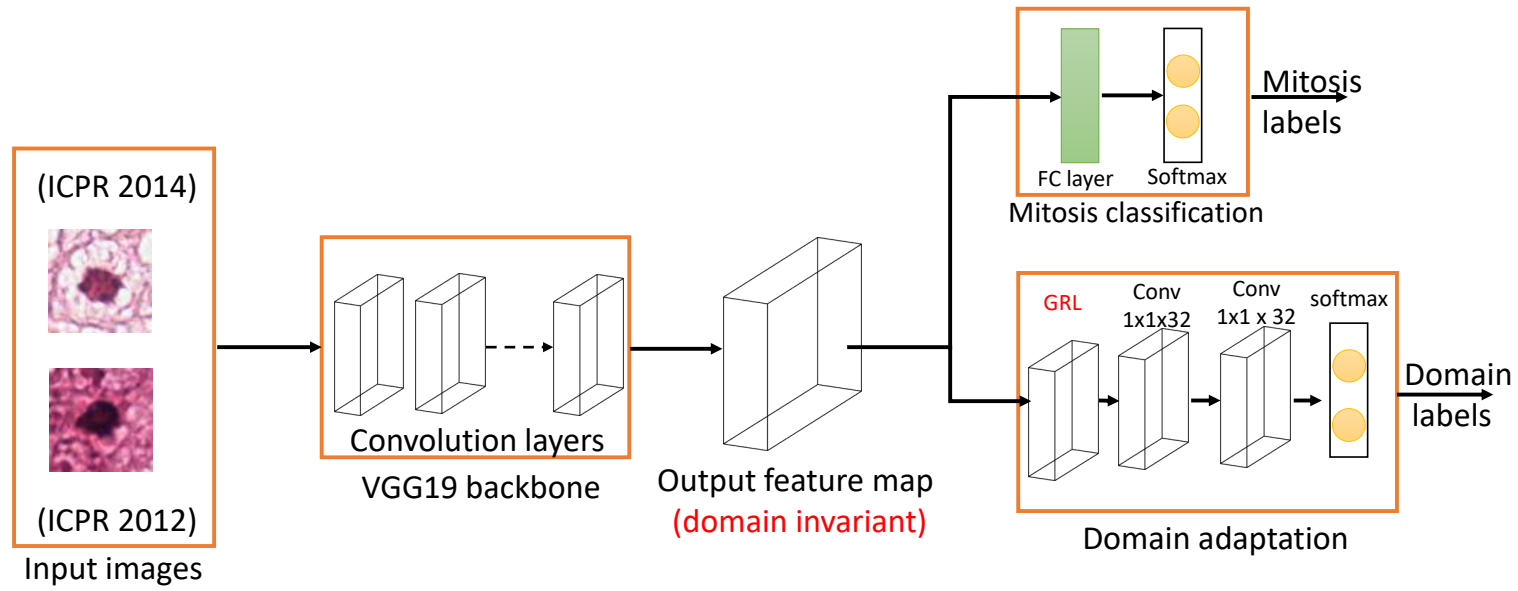


FIGURE 4.3: Domain Adaptive Mitosis Detection Stage: last layer of the VGG-19 network connects to both the mitosis classification branch as well as the domain adaptation branch. This forms the second stage of our dual-stage domain adaptive mitosis detection framework. As in the first stage the gradient reversal layer is used to let the VGG-19 network learn domain invariant features.

4.3 Experiments and Results

4.3.1 Datasets

Experimental evaluation was done on the ICPR 2012 mitosis dataset (Roux et al., 2013) and ICPR 2014 mitosis dataset², depicted in Figure 4.4. Both these datasets contain breast cancer biopsy slides. The slides are stained with standard hematoxylin and eosin (H&E) dyes, and they have been scanned by one of the two slide scanners: Aperio Scanscope XT or Hamamatsu Nanozoomer 2.0-HT.

ICPR 2012 dataset: For this study, we used images scanned by both Aperio Scanscope XT and Hamamatsu Nanozoomer 2.0-HT scanners. The Aperio scanner is referred to as "Scanner A" and the Hamamatsu scanner is referred to as "Scanner H". The dataset has 50 High Power Field (HPF) images sampled from breast cancer histopathology images, at $\times 40$ magnification, for each scanner type. For each scanner, we used 35 HPF images for training and 15 for testing, hence across both scanners there were 70 training and 30 testing images. The images from the Aperio XT scanner have a resolution of 2084×2084 pixels for each field and those from Hamamatsu scanner are 2252×2250 pixels. Across all the images scanned from both Aperio and Hamamatsu scanners, there were 202 mitosis cells in the test set and 552 in the training set. Ground truth masks were available for the mitosis cells in images scanned from both scanners.

ICPR 2014 dataset: The dataset has 1200 High Power Field (HPF) images, sampled from breast cancer histopathology images at $\times 40$ magnification. The corresponding test dataset has 496 image but were not considered for this study since they lack annotations. The 1200 images in the training set have a resolution of 1539×1376 pixels and contain 610 mitosis cells. The training dataset was split into training and validation sets following the same split protocol as mentioned in (Li et al., 2018a), so that we could make a fair comparison. For this study, we considered the images from the from the Aperio XT scanner only. ICPR 2014 dataset is annotated with only centroids of the mitosis cells, hence it can be considered as weakly annotated for segmentation task. For the evaluation, the ICPR 2014 training dataset was split into a training set and a validation set following

²Mitos and Atypia 14 <https://mitos-atypia-14.grand-challenge.org/Dataset/>

the method specified in (Li et al., 2018a; Dodballapur et al., 2019). To determine the ICPR 2014 validation split of the ICPR 2014 training dataset we ran the implementation of (Dodballapur et al., 2019) and ran the validation against many splits until we got the same results as (Dodballapur et al., 2019) and used that validation split for our experiment. So, this makes performing the experiments against that validation split fair and allowed a fair comparison of our results on ICPR 2014 dataset with other state-of-the-art methods.

In the ICPR 2012 dataset, the images scanned using scanner A appear different from those of scanner H due to the difference in scanner resolution as well as in the color output due to the spectral response of their respective sensors and other factors such as the light source and scanning software etc. Furthermore, the ICPR 2014 dataset when compared to the ICPR 2012 dataset has more images with inconsistent staining, tissue folding, and hard negative examples (Li et al., 2018c). These differences in staining techniques and scanning hardware cause a domain shift between the ICPR 2012 dataset and ICPR 2014 dataset as well as within the ICPR 2012 dataset, between scanner A and scanner H images.

4.3.2 Implementation and Results

We conducted two sets of experiments in our study to evaluate mask generation and mitosis detection. For mask generation we performed both qualitative and quantitative evaluation. To evaluate mitosis detection, we performed quantitative evaluation where we trained two domain adaptive models.

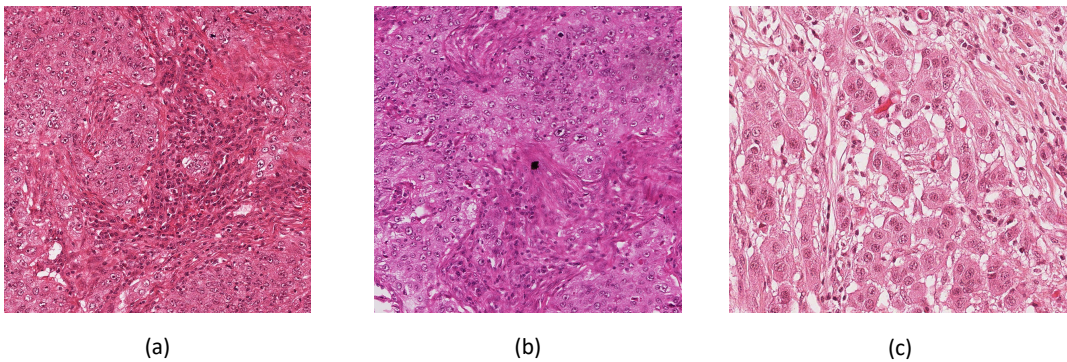


FIGURE 4.4: Sample whole slide images from ICPR 2012 and 2014 datasets: (a) ICPR 2012 using Aperio scanner; and (b) ICPR 2012 dataset using Hamamatsu scanner, (c) ICPR 2014 dataset using Aperio scanner.

- (1) The first model used images of the ICPR 2012 (scanner A) dataset as the source domain and the training split of the ICPR 2014 training dataset as the target domain. This model was tested on ICPR 2014 validation split of the ICPR 2014 training dataset.
- (2) The second model used images of the ICPR 2012 (scanner A) dataset as the source domain and the ICPR 2012 (scanner H) dataset as the target domain for our model training. This model was tested on ICPR 2012 scanner H test dataset.

All experiments were implemented in Python with Keras and TensorFlow libraries and were run on an NVIDIA GeForce GTX Titan dual-GPU machine. We followed the evaluation protocol specified by the ICPR challenge and used precision, recall, and F1-score as evaluation metrics for reporting on mitosis detection accuracy³. A mitosis detection was considered true positive if it was within 20 pixels of the ground-truth centroid pixel for the ICPR 2012 dataset and 32 pixels in the case of ICPR 2014 dataset. This is in accordance with the ICPR contest rules. For reporting on mask generation accuracy, we followed the evaluation methodology used in segmentation and used the Dice coefficient (Dice, 1945), which is a similarity measure of objects or samples.

4.3.2.1 Domain Adaptive Mask Generation Stage

Our first experiment was to demonstrate the advantage of domain adaptation of the first stage of the pipeline to produce robust masks. We trained a Mask R-CNN model without domain adaptation, using the ICPR 2012 dataset (scanner A training set images) only. Next, we trained a domain adaptive Mask R-CNN model using images from the ICPR 2012 dataset (scanner A) as the source domain and images from the ICPR 2012 (scanner H training set) as the target domain. In our domain adaptation training, the images from ICPR 2012 dataset had ground truth labels, i.e. masks and classification labels, and those from the ICPR 2012 (scanner H) dataset did not include ground truth labels. Thus, the training was supervised in the source domain (ICPR 2012 scanner A) and unsupervised in the target domain (ICPR 2012 scanner H). We evaluated both the non-domain adaptive and domain adaptive models using images from ICPR 2012 (scanner H test set). We demonstrated the advantage of domain adaptive training by showing an improvement in

³Challenge Evaluation Metrics <https://mitos-atypia-14.grand-challenge.org/Evaluation-Metrics/>

TABLE 4.1: Results of the dual-stage domain adaptive mitosis detection on ICPR 2012 Scanner H test set comparing the domain adaptive and non-domain adaptive settings. In the domain adaptive setting, the framework was trained using ICPR 2012 scanner A images with supervision, and ICPR 2012 scanner H training images with no supervision. ICPR 2012 (A) refers to scanner A images and ICPR 2012 (A, H) refers to both scanner A and H images. The results show that the domain adaptive model out-performs the non-domain adaptive model.

Model	Training data	Dice	Recall
Non-domain adaptive (non-DA)	ICPR 2012 (A)	1.50	0.27
Domain adaptive (DA)	ICPR 2012 (A, H)	1.56	0.81

the predicted masks as well as the recall rate of the mitosis cell regions. We have also demonstrated the improvement of predicted masks qualitatively.

Table 4.1 provides the results of the mask generation experiment conducted on the ICPR 2012 scanner H dataset. There is about 7% improvement on average in Dice and about 40% improvement in recall. The result validates the improvement in the accuracy of the segmentation masks generated with domain adaptive network from that obtained without domain adaptation. As noted earlier, there is a domain discrepancy between the ICPR 2012 (scanner A) dataset and the ICPR 2012 (scanner H) dataset due to variations in scanner hardware and software. By training our domain adaptive Mask R-CNN stage in an adversarial manner, the network learns features that are intrinsic to mitotic cells such as shape, texture, contours etc., while ignoring artefacts which are introduced by scanner hardware and software such as colour and lighting. Thus, when this network is evaluated on the ICPR 2012 (scanner H), the network is able to predict masks based on features dependent on the appearance of mitosis cells only. We also performed a visual qualitative analysis of the mask generated by the domain adaptive Mask R-CNN and compared it to the non-domain adaptive network. It can be clearly seen that for the examples shown in Figure 4.5, the predicted mask of the domain adaptive model closely resembles the ground truth.

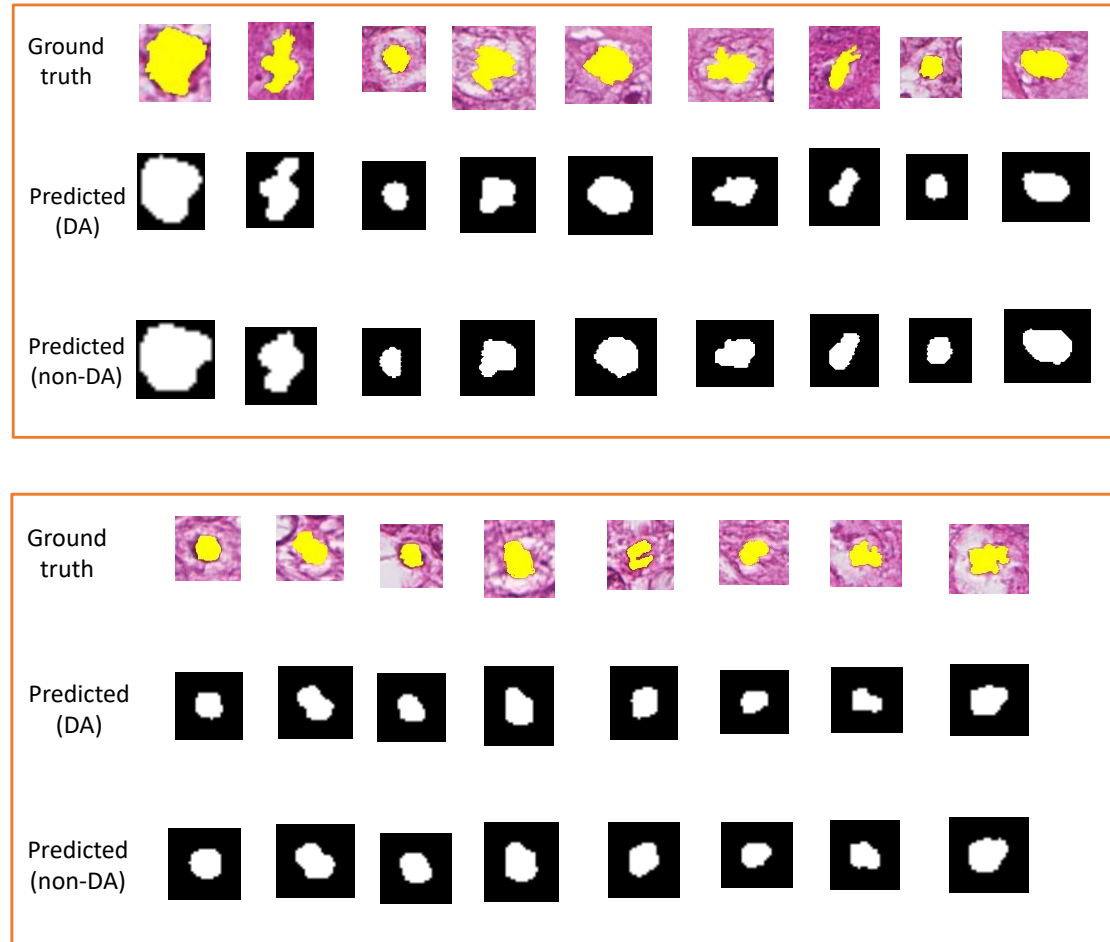


FIGURE 4.5: Examples of masks predicted by the domain adaptive Mask R-CNN (DA) compared against ground-truth images from the ICPR 2012 dataset (scanner H). Visually, it can be seen that masks generated by domain adaptive Mask R-CNN closely resemble ground-truth in both shape and size when compared to non-domain adaptive Mask R-CNN (non-DA).

TABLE 4.2: Results of the dual-stage domain adaptive mitosis detection on the ICPR 2014 validation split (scanner A) and comparison with the state-of-the-art. In the domain adaptive setting, the framework was trained using ICPR 2012 scanner A images with supervision and ICPR 2014 scanner A training images with no supervision. ICPR 2012 (A) refers to scanner A images of ICPR 2012 dataset, and ICPR 2014 (A) refers to scanner A images of ICPR 2014 dataset. In both precision and F1 scores, our domain adaptive method outperforms state-of-the-art even though the other methods have been trained in a supervised setting for ICPR 2014 dataset.

Method	Training data	Precision	Recall	F1
DeepMitosis (Li et al., 2018a)	ICPR 2014 (A)	-	-	0.57
MaskMitosis (Dodballapur et al., 2019)	ICPR 2014 (A)	0.58	0.82	0.68
RCNN (Li et al., 2018c)	ICPR 2014 (A)	0.65	0.66	0.66
Non-domain adaptive	ICPR 2012 (A)	0.55	0.72	0.62
Domain adaptive	ICPR 2012 (A), ICPR 2014 (A)	0.70	0.73	0.71

4.3.2.2 Domain Adaptive Mitosis Detection Stage

The second experiment demonstrates the advantage of domain adaptation of the second stage of the pipeline for the improvement in mitosis detection accuracy. We trained our dual-stage neural network without domain adaptation using the ICPR 2012 dataset (scanner A) only. Next, ICPR 2012 (scanner A) images were used for supervised training of the Mask R-CNN model and the ICPR 2014 images were used for the unsupervised domain adaptation branch only. We then trained the second stage of our pipeline with domain adaptation using predicted masks from first stage for both ICPR 2012 dataset (scanner A) and images from the ICPR 2014 training split. To demonstrate the wider applicability of our domain adaptation framework, we repeated our experiment with ICPR 2012 (scanner A) as the source domain and ICPR 2012 (scanner H) as the target domain. We evaluated the mitosis detection accuracy of the trained model on the ICPR 2014 validation split dataset as well as on ICPR 2012 (scanner H).

Our results for ICPR 2014 validation split in Table 4.2 show that our method outperforms the non-domain adaptive method. This validates our hypothesis that the network learns

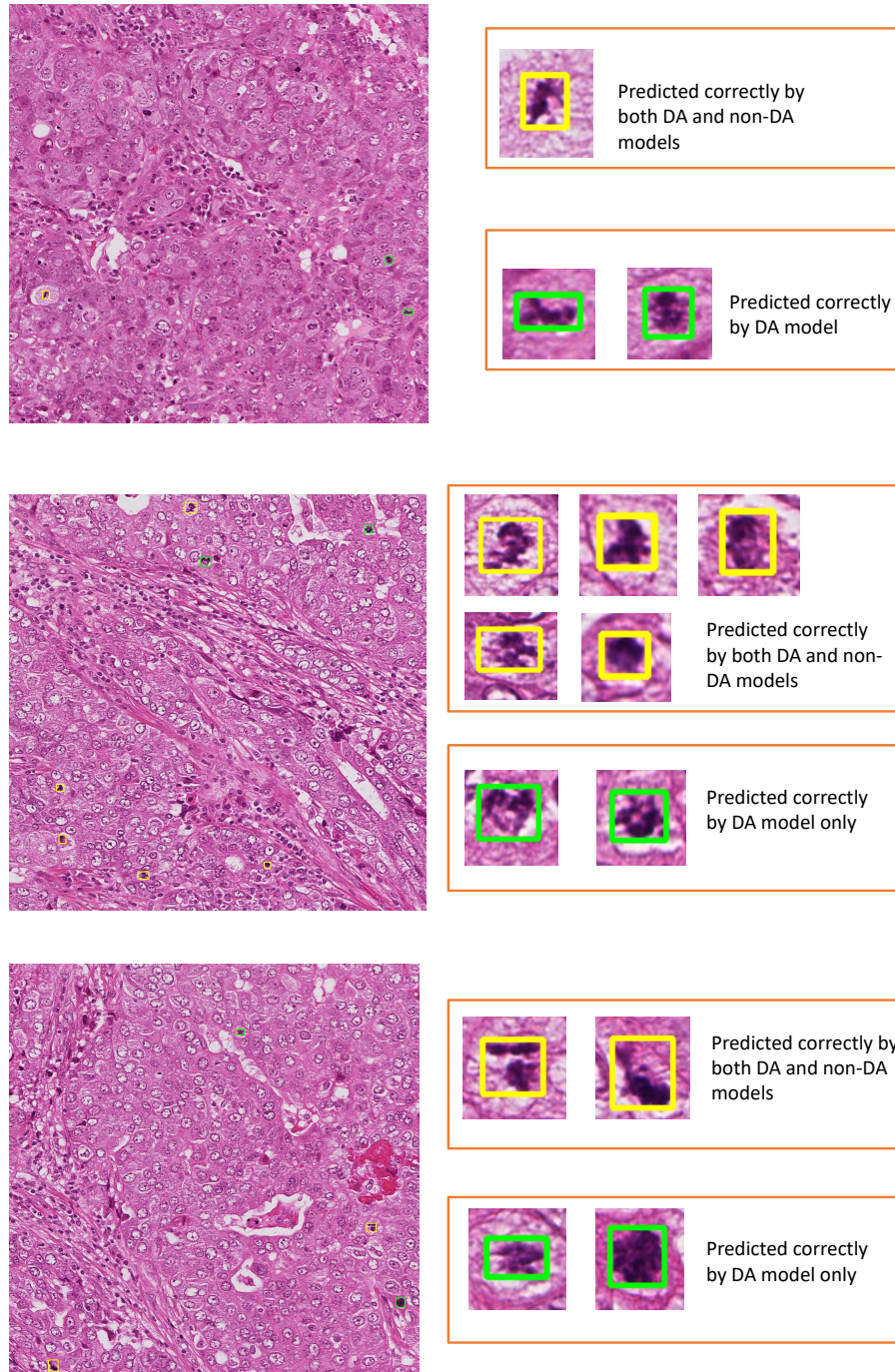


FIGURE 4.6: Examples of the domain adaptive mitosis detection stage (second stage) results on the ICPR 2012 dataset (scanner H) by both domain adaptive (DA) and non-domain adaptive (non-DA) models. When predictions are compared against the ground-truth, the DA model has more correct predictions than the non-DA model.

features which are domain invariant for both the mask prediction stage and the mitosis detection stage while ignoring extraneous details such as colour and lighting variations

TABLE 4.3: Results of the dual-stage domain adaptive mitosis detection on the ICPR 2012 (scanner H) test set, and comparison with the state-of-the-art. We note that the recall rate of the domain adaptive model gets closer to state-of-the-art which has been trained with full supervision on the ICPR 2012 (scanner H) training dataset.

Method	Training data	Precision	Recall	F1
SUTECH (Tashk et al., 2013)	ICPR 2012 (H)	0.82	0.61	0.70
Non-domain adaptive	ICPR 2012 (A)	0.34	0.28	0.31
Domain adaptive	ICPR 2012 (A, H)	0.51	0.55	0.53

(between ICPR 2012 and ICPR 2014 datasets) which do not contribute towards mitosis prediction. Our domain adaptive method also outperforms the state-of-the-art methods (Li et al., 2018c; Dodbballapur et al., 2019; Li et al., 2018a) in both F1-score and precision for the ICPR 2014 validation split. While the state-of-the-art methods were trained and tested on the same domain, which is the ICPR 2014 training split (scanner A) or in other words, (Li et al., 2018a,c; Dodbballapur et al., 2019) do not use any domain adaptation, our method was trained on both ICPR 2012 (with ground truth labels) and ICPR 2014 (without ground truth labels). We argue that our superior performance is because of two factors, namely:

- (1) learning domain invariant features at both stages leads to better model generalization which leads to better network performance at inference (test) time.
- (2) use of mask prediction leads to better localization of mitosis cells.

Compared with (Dodbballapur et al., 2019), we achieve superior performance due to use of domain invariant features. When compared with (Li et al., 2018c,a) both domain invariant features and mask prediction lead to superior results of our method, in other words, our method not only bridges the domain gap between ICRP 2012 and ICPR 2014 datasets but also outperforms methods trained solely on the target domain without mask prediction.

Table 4.3 shows the results of the performance of domain and non-domain adaptive models on ICPR 2012 (scanner H) dataset. To the best of our knowledge, this is the first experiment conducted in literature, where images from one scanner (scanner A) were used

for training and tested on images from another scanner (scanner H). Our domain adaptive model achieves a 70% improvement in F1 score when compared to the non-domain adaptive model. Figure 4.6 compares the performance of our domain adaptive method with our non-domain adaptive method when tested on ICPR 2012 scanner H dataset. From the examples illustrated in Figure 4.6, it can be clearly seen that the domain adaptive method comes up with more correct predictions when compared to the non-domain adaptive method. However, our domain adaptive method does not perform as well as (Tashk et al., 2013) which is trained and tested on ICPR 2012 (scanner H). In other words, our domain adaptive training does not completely bridge the domain gap that exists between ICPR 2012 scanner A and scanner H datasets. We argue that there is a large colour variation between the images of scanners A and H and hence a larger domain gap, when compared to images within the same scanner. As a future work, we hypothesize that training our domain adaptive model with larger datasets may help overcome this domain gap.

4.4 Summary

For mitosis cell detection, the domain shift is mainly due to different staining conditions. Supervised domain adaptive solutions have provided some success to address this problem but requires annotated datasets. In this chapter, we have proposed a method of domain adaptive mitosis detection by generating masks automatically using unsupervised domain adaptation. We show that our method predicts robust and more accurate masks for weakly annotated datasets such as ICPR 2014 as well as the ICPR 2012 (scanner H) dataset. We have also shown that the accuracy of the mitosis detection can be further improved by applying a similar unsupervised domain adaptation in the second refinement stage of our pipeline. Learning domain invariant features at instance and image level leads to better network generalization and improved mitosis classification accuracy. While we have demonstrated the advantage of our framework for mitosis detection, this dual stage detection method can be adapted to other applications such as pedestrian or vehicle detection, where objects need to be recognized and localized in different viewpoints or conditions.

Cross Domain Exemplars for Cell Counting

Supervised learning provides good results, as the data is labelled. However, obtaining labelled data is often a very tedious operation and error-prone, especially for instance-based segmentation. Hence, there are growing number of methods where the data is only partially labelled. The unlabelled data could be incomplete, inexact, or inaccurate. In the cross-domain datasets, the target data is completely unlabelled. Though there are many methods, as discussed above in the Section 1.2.1, to handle weakly supervised data, it still has challenges to overcome. In inexact annotations, including dot annotations, the dot may appear anywhere in the cell. This poses a challenge in accurate localization of the cells. Additionally, data with noisy background and overlapping objects, also pose a challenge in training a robust model.

Cell counting plays a significant role in biological studies and medical diagnosis. To achieve automatic counting of cells in an image, there is a need to create annotated datasets, which is not always practical. Furthermore, the quality of the exemplars plays a key role in automatic cell counting. In this chapter, we introduce a novel approach of weakly supervised learning, by automatically selecting good quality exemplars obtained from cross-domain training. This chapter is built on the fully supervised methods introduced in the first part of the thesis. We exploit cell structure similarity to achieve cross-domain knowledge transfer. Firstly, two domain adaptive Faster R-CNN networks are trained on RGB microscopy images, and background subtracted images, respectively. The candidate exemplars obtained from each trained network are merged based on their confidence scores, and then through K -means clustering approach, good quality exemplars are selected based on their appearance variability. A density map prediction network is trained using a small fraction of the chosen exemplars to count the number of cells. In

our experiments on the MBM, DCC and VGG datasets, we achieved close to state-of-the-art results, using far less training samples. The content of this chapter has been accepted at IEEE International Symposium on Biomedical Imaging, ISBI 2024¹ and is currently not available online.

5.1 Introduction

In real-world applications, there is a need to count a large number of objects in images, such as cells, wild animals, persons etc. Microscopy cell counting provides enormous advantage for medical diagnosis by analyzing the cells in biological studies (Mukherjee et al., 2017). However, manual counting of these cells is both tedious and error-prone as there are a large number of these cells. Further, the cells are often overlapping with their edges diffused. In addition, the images in which the cells appear are stained with different backgrounds. The cells also vary in structure depending on the cell types. Finally, annotating cells can be expensive and time-consuming. All these factors pose challenges to counting cells in images automatically using machine learning methods.

Exemplars are distinct objects within an image, like cells, animals, or cars, which are selected as suitable representatives for counting them in the image (Ranjan et al., 2021). These exemplars should be of good quality to get an accurate count. In the context of cell counting, ideal cell exemplars are characterized by their distinctness, non-overlapping cells, uncluttered background, good shape and texture features as illustrated in Figure 5.2 . Manual selection of exemplars is a subjective process as it depends on a human in the loop. Further, in applications where cell counting needs to be fully automatic, human annotation of exemplars will not be possible. Hence, it is desirable to automatically choose good quality exemplars.

In the domain of microscopy cell counting, automatic exemplar selection will certainly provide an advantage as there are a large number of cells to count and within an image the appearance of the cells varies subtly. This poses a challenge in defining what makes a good exemplar for counting, as it relies on the skill of the person to define the exemplars, or the system needs to be robust to bad exemplars. Thus, automatic exemplar

¹<https://biomedicalimaging.org/2024/>

generation, apart from removing the need for manual annotation, also allows for selection of good exemplars automatically in an objective manner which would be advantageous. In microscopy cells, even though the cell domains are different, structurally, at a gross structural level certain types of cells look similar. For example, DCC and MBM cells are oval or circular in shape (Paul Cohen et al., 2017), (Marsden et al., 2018). Hence, it will be advantageous to disentangle the cell appearance, and transfer the knowledge from a source domain where an exemplar is defined, to a target domain where an exemplar is not available. Such a domain transfer would not only make exemplar generation automatic but also enable accurate counting of all cells in the image based on a few good quality exemplars. The following section summarizes the related work in this field.

5.1.1 Related Work

In the literature, object counting methods can be categorized into detection-based methods (Chattopadhyay et al., 2017), (Laradji et al., 2018) and regression-based methods (Cholakkal et al., 2019), (Liu et al., 2019c). In the detection-based methods, objects are detected, and the detections are counted. In the regression-based methods, a density map is predicted and the count of the number of objects is determined by integrating the pixel values of the density map. The sections below describe both the cell detection- and regression-based approaches.

5.1.1.1 Detection Approaches

For automatic microscopy cell counting, one of the earlier studies uses shape and concavity detection within a cluster of cells (Kothari et al., 2009), segment the cluster into individual nuclei, to count them. Similarly, (Jung and Kim, 2010) segments clustered nuclei of breast and cervical cells, by solving an optimization problem within the watershed framework. Furthermore, there have been counting approaches based on cell curvature (Zhang et al., 2012) and distance transforms (Jung et al., 2010) to determine clusters of cells. Hough Transform has been used to detect and isolate red blood cells and count them for detection of cancers such as leukaemia (Maitra et al., 2012). These techniques, however, are not very robust when cells are varied and fused and thus have had only a moderate success with counting overlapping cells. They especially performed poorly as

the cell density increased. An early study (Arteta et al., 2012) for cell counting using machine learning, used maximally stable regions to train a structured Support Vector Machine (SVM), to get the cell count. Following on from the advancement of deep learning algorithms, a number of CNN approaches for cell counting have been proposed. The deep learning based study (Chen and Ched'Hotel, 2014), uses sparse colour unmixing to extract the color channel corresponding to the cell membrane and formulates an object detection problem using a DNN, on that image channel to detect and count cells. Another deep learning based study (Liu and Yang, 2017), uses DNN to score image patches around the cell detections based on their proximity to the true cell centers.

Cell counting is performed by constructing graph of cell patches based on their scores and finding maximum-weight independent sets within the graph. Based on the recent success of region proposal networks, the study (Yusro et al., 2023), trains a Faster R-CNN network to count blood cells. Instance segmentation methods using both Mask R-CNN and U-Net have been proposed for cell detection and counting (Paulauskaite-Taraseviciene et al., 2019), where the authors demonstrate that U-Net performs as well as Mask R-CNN despite being a simpler network. Residual U-Net which is an extension of the U-Net architecture, widely used within the medical imaging community, has been applied to cell counting (Rad et al., 2018), through multi-scale training. The detection-based methods localize the cells well and achieve good counting accuracy. However, all detection-based approaches require fully annotated datasets, with either bounding box or mask annotation for the cells. Creating and obtaining fully annotated datasets is laborious and often impractical.

5.1.1.2 Regression Approaches

The regression-based approaches rely on weakly annotated datasets containing only dot annotations for cells. The study (Arteta et al., 2014), uses an interactive density counting approach, where the network predicts object density by training on images with dot annotations. An interactive visualizing tool is presented to the user to inspect the predictions and correct them, thus creating more annotated examples in an interactive manner. In another approach (Akram et al., 2016), a cascade architecture with non-max suppression is used for cell detection and counting. The study (Xie et al., 2018), trains a fully convolutional regression network using synthetic datasets to predict density maps, and

applies it successfully to count cells in real cell datasets. Rather than predicting a density map, Count-ception (Paul Cohen et al., 2017) uses a DNN from the Inception family, to obtain redundant counts for each pixel location within an image, by considering small receptive fields. The count at every image pixel is averaged from the redundant counts, and the overall count is obtained by integrating the pixel values for the entire image. The study (Jiang and Yu, 2020), proposes a multi-column, and multi-resolution extension to U-Net network, by replacing the convolutional layers with residual blocks, to count cells of varying sizes. Sau-Net (Guo et al., 2021) uses an extension for batch normalisation and attention-based U-Net, to handle small datasets, for cell counting. Regression based approaches, have the advantage of using weak annotations, but the methods cannot provide precise locations of objects, and hence suffer from poor counting accuracy.

5.1.1.3 Weakly Supervised Counting

The weakly supervised counting methods use a reduced number of cells for training. These reduced number of cells are called exemplars which are representative cells in the image. Annotation cell counting (Wang and Yin, 2021) uses a semi-supervised technique for cell counting which limits the training data to some extent. This method adapts active learning to use exemplars selectively based on uncertainty and diversity. A cross-domain solution study (Wang, 2023), uses a synthetic dataset as a source domain dataset and a partially annotated target domain dataset to train a network in a cross-domain setting. The network learns domain agnostic features, which is fine-tuned using synthetic images, generated for the target domain using a GAN. However, this method relies on generative adversarial networks which are difficult to train and require a fully annotated subset of training images to generate synthetic images. FamNet (Ranjan et al., 2021) proposes a solution which requires a few exemplars for each image, which is typically provided through human annotation. It uses a correlation of the image features and exemplar features to predict a density map, which is used for cell counting. Repetitive Region Proposal Network (RepRPN) (Ranjan and Nguyen, 2022) uses region proposal network to find candidate objects and selects the appropriate candidates with highest repetition within an image, as exemplars. In this manner, exemplars are automatically selected and used to predict density maps for counting. Finally, the zero-shot counting approach (Xu et al., 2023), uses quality of density maps to determine exemplars, and uses these exemplars for

counting in a subsequent stage of the pipeline. However, none of the existing techniques provide a reliable and robust method to choose good quality exemplars automatically for cell counting.

To overcome the lack of quality exemplar selection for cell counting, we propose a method of transferring the knowledge from a source domain, where exemplars are defined, to a target domain, where exemplars are not defined, by exploiting the structural and textural similarity of cells between domains, to automatically generate good exemplars in the target domain for cell counting. This makes the task of cell counting completely automatic in the target domain by obviating the need for a human to provide exemplar annotations. In addition to enhancing cell counting accuracy, the exemplar selection method makes the density map prediction training more efficient, by considering a smaller subset of higher-quality training data.

5.1.2 Contributions

In this chapter, building on the supervised methods of first part of thesis, we address the challenge of weak and incomplete dataset annotation as outlined in Section 1.5.4, where the weakly annotated datasets do not have complete annotations required for good localization. This, in turn, gives rise to poor quality exemplars and has an impact on cell counting, as the cells are weakly annotated and are varied and densely packed. Hence, our contributions to address these challenges are summarized follows:

- (1) We propose a method to automatically generate exemplars with their respective confidence scores by training two domain adaptive region proposal networks;
- (2) We design a clustering technique to further refine the selection of exemplars to account for subtle variations in the cell appearance;
- (3) Based on the automatically generated exemplars, a network is then trained to predict a density map which in turn is used to obtain a count of the number of cells by integrating the pixel level density values;
- (4) Our method provides close to state-of-the-art results on the VGG (Lempitsky and Zisserman, 2010), DCC (Marsden et al., 2018), and MBM (Paul Cohen et al., 2017) cell datasets whilst using fewer training examples when compared to the other proposed methods in the literature.

5.2 Method

Our automatic cell counting method has three stages. The first stage is the domain adaptive exemplar generation stage, which detects cells for the input RGB microscopy images and pre-processed microscopy images using a region proposal network trained in a cross-domain setting. In the second stage, the candidate exemplars and their respective confidence scores are clustered using K -means clustering. For clustering, we generate appearance-based features for the exemplars using a ResNet-50 network (He et al., 2016) pre-trained on ImageNet. Representative exemplars from the clusters are chosen for cell counting using both cluster membership and confidence score. In the third stage, we train a network to generate the density map based on the input images and their respective exemplars.

5.2.1 Exemplar Generation

In this stage, we used a domain adaptive region proposal network such as Faster R-CNN (Girshick, 2015) network to automatically detect cells in the image. This process is illustrated in Figure 5.1 (a). The network uses ResNet-50 as the backbone and is pre-trained on ImageNet. We added a domain adaptive branch with a gradient reversal layer, which helped the overall network learn domain invariant features, in our case, the structural similarities of the cells between source and target domains. This domain adaptive network was trained with labeled source domain dataset and unlabeled target domain dataset. In this manner, we trained two domain adaptive Faster R-CNN networks. The first network had the source and the target from the original RGB microscopy datasets. The second network was trained using the RGB images of the source dataset and pre-processed images of the target dataset where the target domain images were pre-processed using an image pre-processing step described as follows.

In the pre-processing step, a thresholding approach was used to obtain background subtracted microscopy images. Firstly, the image was blurred using a Gaussian kernel and then binned to reduce the resolution. For each dataset that we evaluated on, we determined a suitable threshold empirically. The result of applying the pre-processing step is illustrated in Figure 5.3 where the first row has examples of the original images from

the datasets we used, and the second row has the corresponding background subtracted images.

The candidate detections were then merged based on their respective confidence scores as shown in Figure 5.5. The detections from the two networks were filtered based on their average confidence score whenever they intersected. For those detections that did not intersect, the confidence score was halved. The combined list of detections was then filtered based on an empirically determined threshold.

5.2.2 Exemplar Selection

Using the candidate exemplars determined after confidence-based merging, we computed appearance-based features for the candidate exemplars and applied K -means clustering on the features. We used a ResNet-50 network (pre-trained on ImageNet) to extract features for each of the candidate exemplars. The individual clusters capture subtle variations in the cell appearance. From each cluster, based on the confidence scores, best exemplars were chosen, to make up an optimal set of exemplars for density map generation. The steps of this process are illustrated in Figure 5.1 (b), and the result of the clustering step is illustrated in Figure 5.4. The selected exemplars were used to train a network to predict density maps.

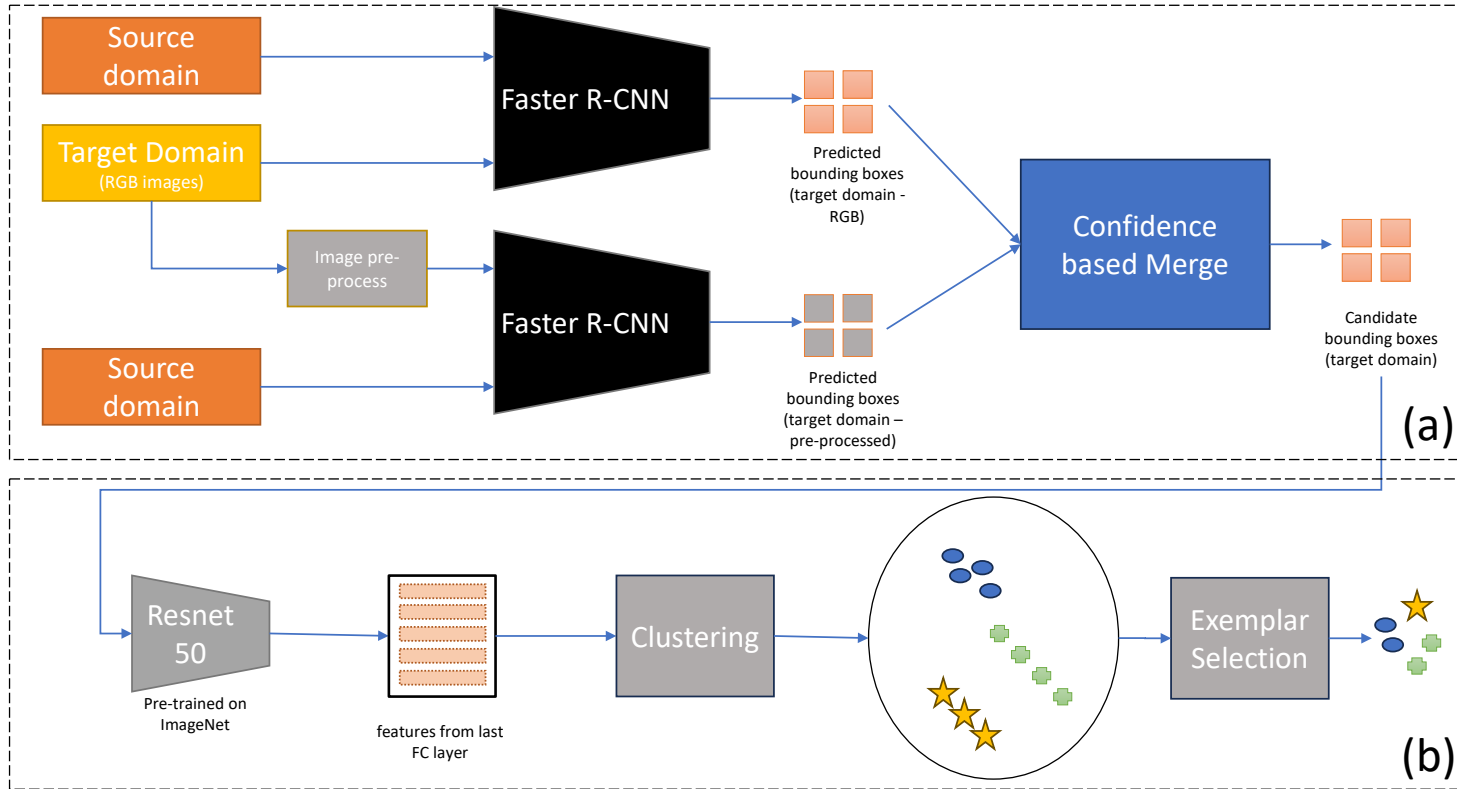


FIGURE 5.1: Cross-domain exemplar generation and selection framework: (a) Two Faster R-CNN networks are trained with RGB, and background subtracted images, respectively. The detections which are bounding boxes predicted by the Faster R-CNN network are merged based on overlap and confidence scores; (b) Candidate bounding boxes from (a) are clustered based on features extracted from a ResNet-50 backbone and exemplars are selected from cluster centroids based on their respective confidence scores.

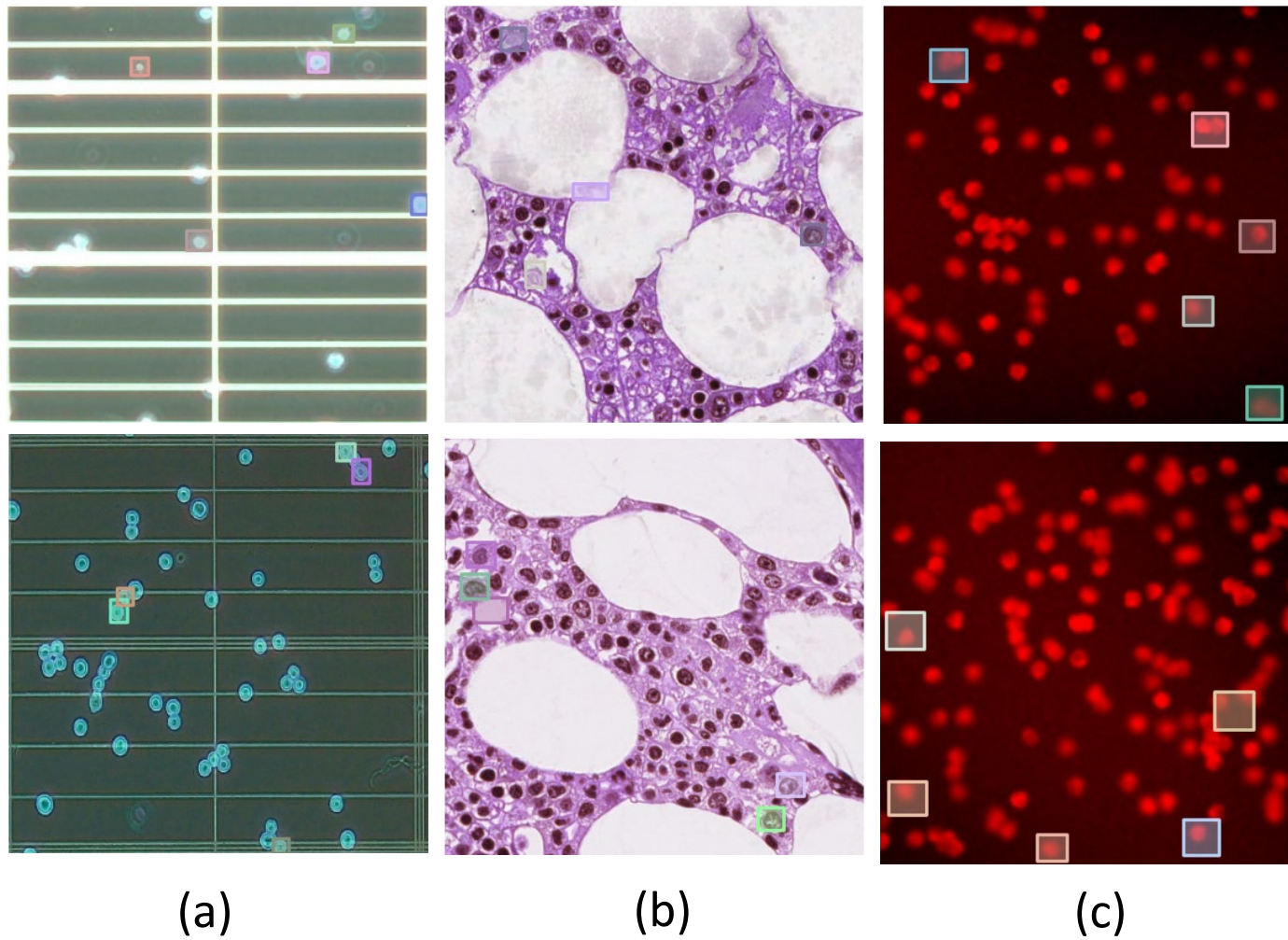


FIGURE 5.2: Illustrative exemplars selected from images of three different datasets we used in our evaluation: (a) DCC dataset, (b) MBM dataset; and (c) VGG dataset. For each dataset, we illustrate with bounding boxes, the exemplars that have been chosen. Note that for the VGG dataset, for clarity, we have shown the cells in red colour instead of blue.

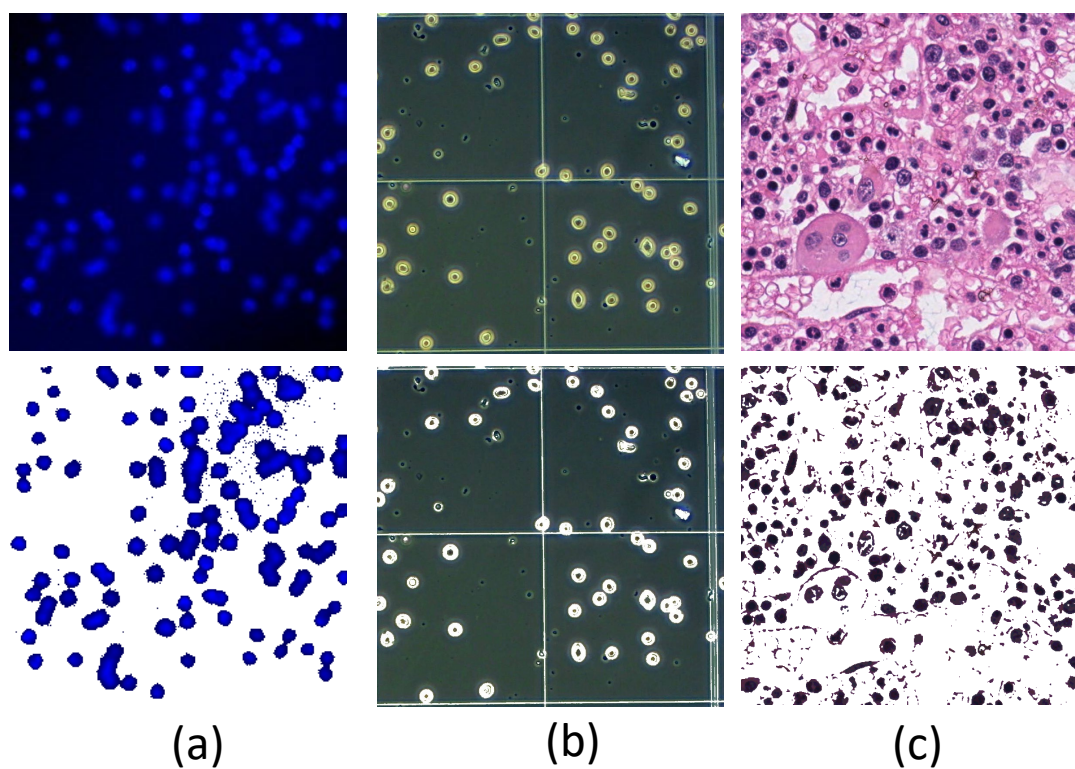
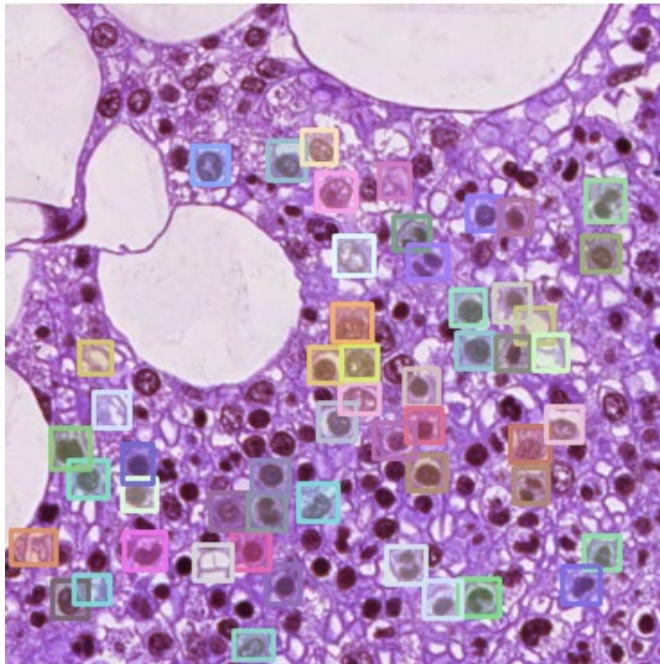
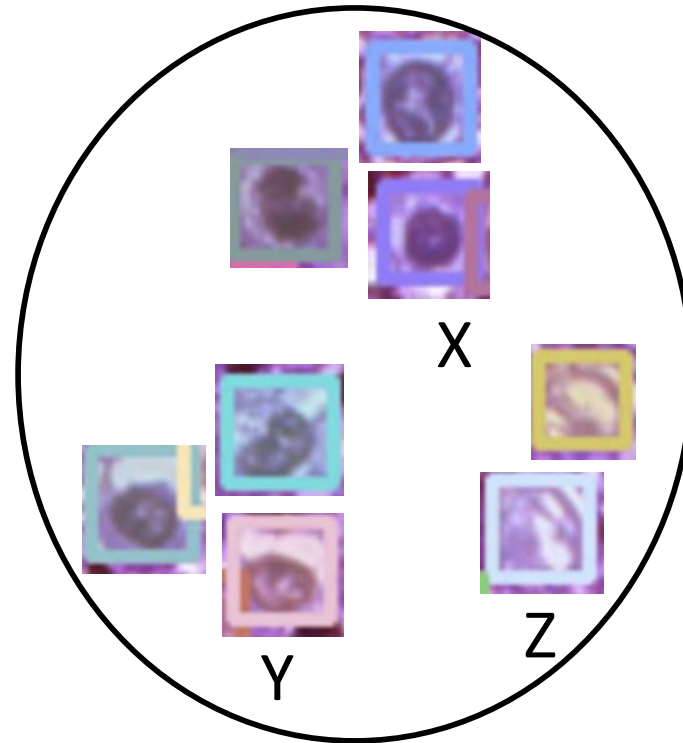


FIGURE 5.3: Result of background subtraction step applied to cell images from the three datasets we used for our evaluation. The first row contains the original RGB images, and the second row has the corresponding background subtracted images: (a) VGG dataset; (b) DCC dataset; (c) MBM dataset.



(a)



(b)

FIGURE 5.4: Results of K -means clustering of candidate exemplars to account for variation in cell appearance. (a) Image with candidate exemplars determined by merging bounding box predictions from the two Faster R-CNN networks. (b) Three clusters X, Y, Z as determined by K -means using ResNet-50 features extracted from image patches of the candidate exemplars. The clusters are representative of the variations in the cells.

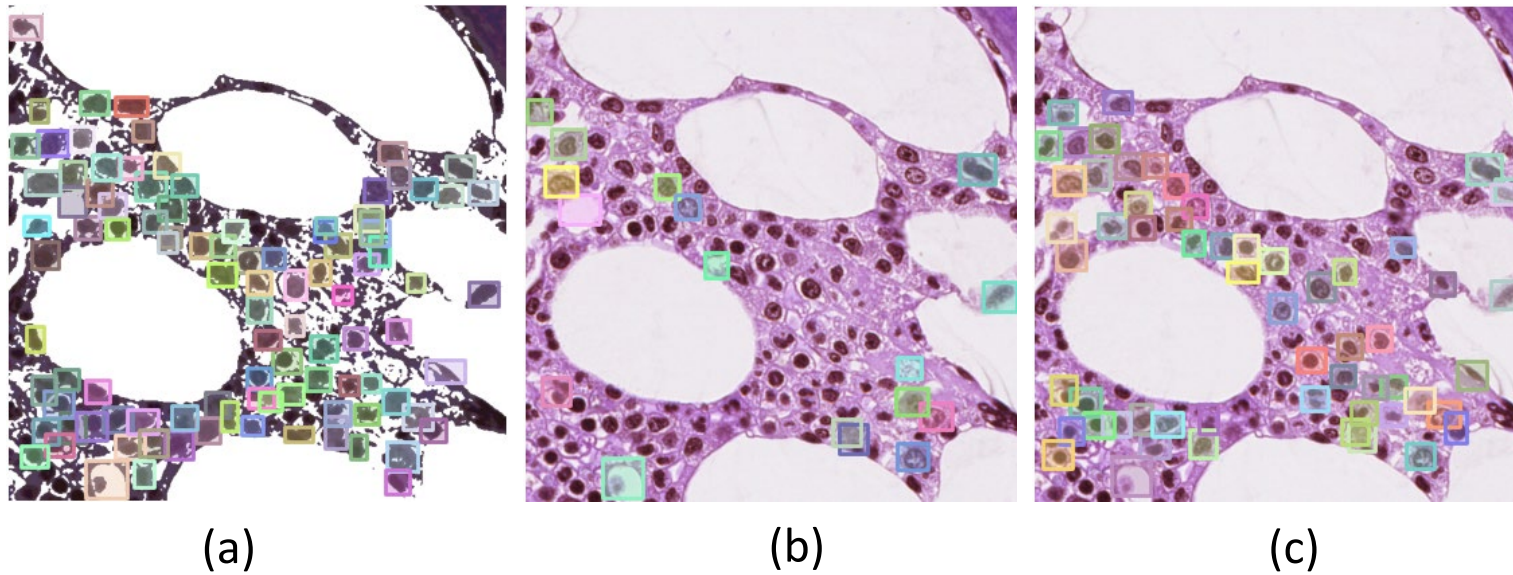


FIGURE 5.5: Results of determining candidate detections through confidence-based merging: (a) Detections obtained from bounding box predictions on the background subtracted image; (b) Detections obtained from bounding box predictions on the RGB image; (c) Merged candidates, combining detections from RGB, and background subtracted images.

5.2.3 Density Map Generation

The exemplars obtained as described in Section 5.2.1, were used for cell counting. We used a regression-based density map generation approach similar to the studies (Ranjan et al., 2021), (Xu et al., 2023). A counting deep network module, comprising a feature generator, and density map predictor, was used. The exemplars were augmented using different augmentation techniques such as rotation, flipping, and scaling.

A ResNet-50 (He et al., 2016) network pre-trained on ImageNet was used to generate appearance-based features based on the image patches of the exemplars. The features were extracted from the third and fourth convolution layers. Then, an ROI pooling was performed to get scale invariant features, at different scales. The features thus obtained were correlated with the image features, a correlation matrix was determined. Correlation matrices obtained for all augmented exemplars were averaged to get the final correlation matrix.

We used weak dot annotations of cells to generate ground-truth density maps to train the density map prediction network. A Gaussian filter with sigma value of 3.5 was used to generate the density maps. The averaged correlation matrix from the previous step, and the ground truth density map were used to generate the density map predictions using a network with five convolution blocks. The network followed the encoder-decoder architecture, with the last layer being a single channel output, and MSE (Mean square error) loss function.

5.3 Experiments and Results

5.3.1 Datasets

We performed our evaluation on three popular cell datasets VGG, DCC and MBM as shown in Figure 5.6. In all the three datasets, only dot annotations were provided. The dot annotations were used to generate ground-truth density map with a Gaussian kernel of sigma value set to 3.

VGG Dataset: The synthetic dataset (Lempitsky and Zisserman, 2010) has 200 images

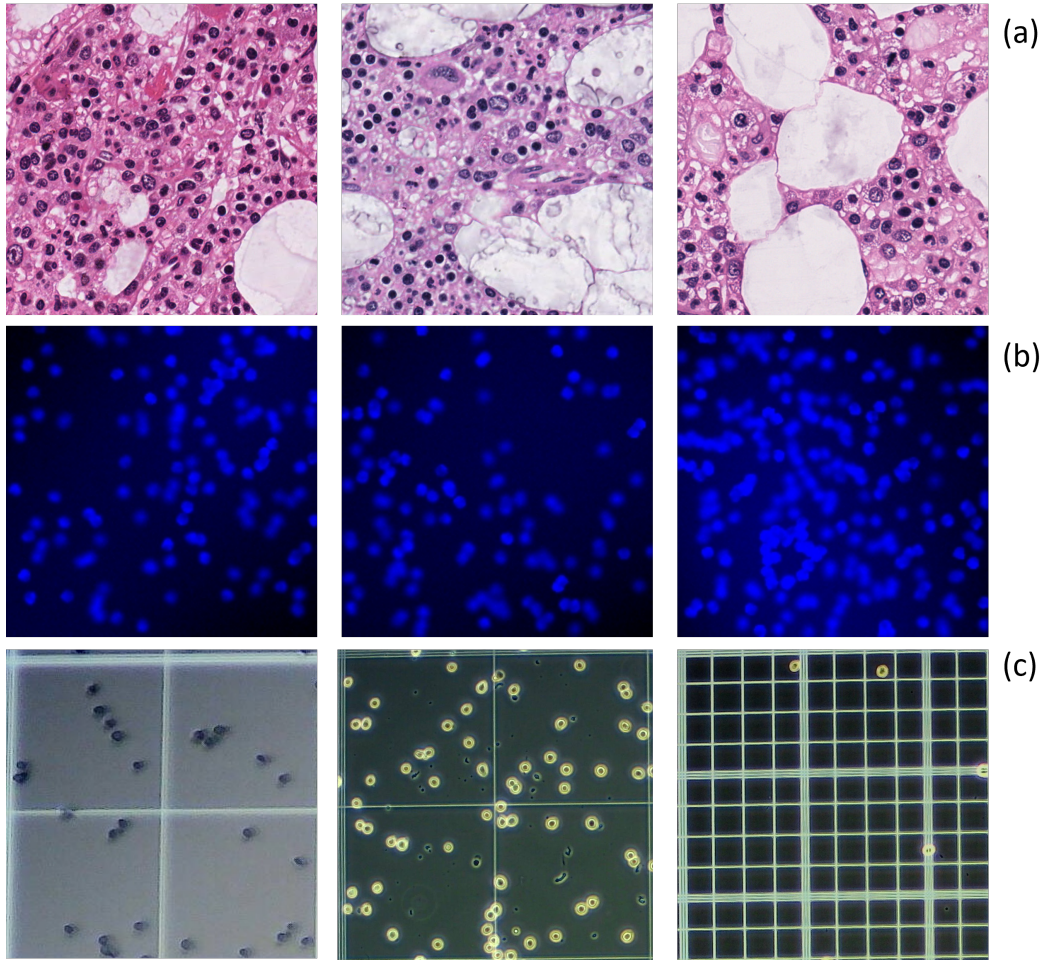


FIGURE 5.6: Datasets used in the experiments: (a) MBM Dataset; (b) VGG Synthetic Dataset; (c) DCC Dataset.

with a resolution of 256×256 . In each image there are 174 ± 64 cells. There are 50 images in the training set and the rest in the test set. The data set mimics bacterial cells from fluorescence-light microscopy at various focal distances.

MBM Dataset: We used the MBM dataset with modifications as described in (Paul Cohen et al., 2017) having 44 images with 126 ± 33 cells per image. The modified MBM dataset was obtained from the original MBM dataset of 1200×1200 resolution proposed by (Kainz et al., 2015). As described in (Paul Cohen et al., 2017), 30 images were used for training and the rest for testing. This dataset contains real images of human bone marrow with various cell types stained in blue.

DCC dataset: There are 100 training images and 76 testing images in the Dublin cell counting dataset (Marsden et al., 2018). The image sizes vary from 306×322 to 798×788 . Each image has 34 ± 22 cells.

5.3.2 Implementation and Results

In training the first Faster R-CNN network, original source, and target domain RGB microscopy images were used, and for the second Faster R-CNN network, original source images, and pre-processed target domain images were used. The network used binary cross entropy loss and bounding-box regression loss. Each network was trained for 200 epochs with stochastic gradient optimization. The learning rate was set to 0.002. The images were resized to 600×600 during both training and inference phases. The outputs of the model were the detection candidates from the two networks and their respective confidence scores. The detections (candidate exemplars) for each image obtained using the original RGB image as well as the background-subtracted image were merged and sorted based on the confidence scores. The top "n" candidates of the detections were selected, where "n" was determined by setting a confidence threshold of 0.6. This confidence threshold was empirically determined and applied to all our experiments. Since only weak annotations were available for the datasets, full supervision for source domain was achieved using randomly generated bounding boxes based on the weak dot annotations for the cells.

The original RGB detections thus obtained from the exemplar generation step, were fed into the ImageNet pre-trained ResNet-50 model. Appearance-based features were extracted from the last fully connected layer for all the exemplar candidate images. K -means clustering was performed and the centroids from each of the clusters were chosen as the final exemplars for obtaining density maps.

For choosing the number of exemplars during training and inference, we performed several experiments with each of our datasets by using from 1 to 5 exemplars during the training step and used it for density map prediction. We noticed that while the accuracy increased with more good quality exemplars, 5 exemplars were sufficient to give a good counting performance for VGG and MBM and 3 exemplars were enough for DCC. For all our experiments, we used the same number of exemplars during both training and inference for density map prediction. The density map generator was trained using features obtained from a ResNet-50 network pre-trained on ImageNet similar to (Ranjan et al., 2021). The image was resized to 384×384 and optimization was done using Adam. The learning rate was set to 10^{-5} . The features were obtained from the 3^{rd} and the 4^{th} block of

TABLE 5.1: Results of cross-domain automatic cell counting in different settings. For each experiment, we note the target and source datasets, with the source indicated in brackets. E.g. DCC(VGG) indicates that DCC is the target and VGG is the source dataset. "Original" indicates cross-domain training done with RGB microscopy images, "Pre-processed" with background subtracted images and "merged" which uses a combination of original RGB and background-subtracted images.

Method	DCC(VGG)		MBM(VGG)		VGG(MBM)	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
Original	5.9	8.2	22.5	30.4	12.6	15.9
Pre-processed	5.4	7.7	16.0	17.7	9.9	12.4
Merged	4.8	6.7	9.53	11.6	6.1	8.7

the convolution layers and combined to obtain a multi-scale feature vector. This was then used to train the density map prediction network which consisted of five convolution layers and three upscaling layers to produce a density map image with the same dimensions as the input image.

We implemented all our experiments in Python with Keras and TensorFlow libraries and were run on an NVIDIA GeForce GTX Titan dual-GPU machine. As in literature, we used MAE (mean absolute error) and RMSE (root mean square error) (Ranjan et al., 2021), (Chai and Draxler, 2014) as the evaluation metric.

We performed three sets of experiments for automatic cell counting for all the datasets:

- (1) using original RGB microscopy images;
- (2) using background subtracted microscopy images;
- (3) combination of both (1) and (2) with merged exemplars.

Table 5.1 shows the results of our experiments for all datasets in all three settings. We note that in all the three datasets, the best performance was achieved when exemplars from original RGB image and background subtracted image were merged. The second-best result was achieved using exemplars from background-subtracted image only. The least accuracy was obtained when using exemplars from the original RGB microscopy

TABLE 5.2: Results of cross-domain automatic cell counting compared with state-of-the-art methods. The "Cells" column indicates the approximate number of cells used at training time which are human annotated (in state-of-the-art). For our method, the cells column indicates the number of suggested exemplars used for density map generation.

Method	DCC		MBM		VGG	
	MAE	Cells	MAE	Cells	MAE	Cells
Ours (Merged)	4.8	300	9.53	150	6.1	250
Sau-Net (Guo et al., 2021)	3.0	3400	5.7	1890	2.6	11136
Ann Cell cnt (Wang and Yin, 2021)	5.1	340	9.0	378	6.2	870
CD disentangled (Wang, 2023)	6.3	238	19.1	252	—	—

image only. When we used both the background-subtracted and the original images for generating exemplars, better quality exemplars were chosen based on both the texture and shape characteristics. Further, when we used only the pre-processed images for exemplar generation, only the shape of the cells was considered, resulting in lower accuracy compared to the combination as used in experiment (3). Finally, when only the original RGB image was used for exemplar generation, the shape information was not given prominence and the noisy background of the image introduced some false positive detections. We also note that in the MBM dataset there were false positive cells. While some of these cells were chosen as exemplars in experiments (1) and (2) but in (3) because of confidence based merging, false positives were either reduced or removed.

In Table 5.2 we compare our method to the state-of-the-art methods Annotation Cell Count (Wang and Yin, 2021) and Cross-Domain Disentangled Cell counting (Wang, 2023). We note that all state-of-the-art methods require some form of human annotation. To make the comparison fair, we have compared our approach to state-of-the-art approaches which use at least as many cells (human annotated) as the generated exemplars used in our method. As can be seen from Table 5.1, we either closely match or improve on the current state-of-the-art. For completeness, we have included Sau-Net, which requires full annotation to achieve state-of-the-art results.

5.3.3 Discussion

In Chapter 3, we studied mitosis cell detection using a two-stage cascaded detection method based on Mask R-CNN (He et al., 2017) and extended this approach in Chapter 4, to train two-stage detection framework based on Mask R-CNN, and evaluated it in a cross-domain setting. In this cell counting study proposed in this chapter, we use Faster R-CNN instead of Mask R-CNN due to unavailability of pixel-level masks in any of the datasets we evaluated on. While there have been several studies in the recent years focussing on Mask R-CNN for cell counting (Loh et al., 2021), (Dhieb et al., 2019), (Naets et al., 2021), the proposed approaches require pixel-level annotations for cells. Creating pixel-level annotations to train a Mask R-CNN network becomes prohibitively expensive especially for datasets with large number of cells. Furthermore, these approaches perform poorly when cells are overlapping or when cells are clumped. Additionally, human error in annotating overlapping cells affects the performance.

Recently, the Segment Anything model (Kirillov et al., 2023), trained on large datasets for segmentation task, has gained a lot of attention. Based on the success of the Segment anything model, its application to medical images is now being studied (Mazurowski et al., 2023). As a future research topic, we propose to explore the combination of Segment anything model and Mask R-CNN to generate good quality exemplars for pixel counting, by predicting pixel level masks rather than bounding boxes. We hypothesize that this approach will yield better localization of exemplars, this leading to better cell counting accuracy.

5.4 Summary

In this chapter, we have proposed cell counting with domain adaptive automatic exemplar selection. Our approach uses three steps. In the first step, two domain adaptive networks with RGB microscopy image and pre-processed images as inputs are used to generate exemplars. The candidate detections from the two networks are merged and filtered based on their confidence scores. Then, the detections are clustered from which the exemplars are chosen. Density map generation is used for cell counting using a small fraction of these exemplars. Experiments conducted on MBM, VGG and DCC datasets show that exemplars are of good quality and produce close to state-of-the-art results.

Spatio-Temporal Region-Specific Features for Person Re-Identification

Domain-shift problem occurs when models trained with dataset in source domain are used in another domain, namely the target domain. Such models do not perform well due to different feature distributions between source and target domain. In fully supervised methods, labelled datasets from all domains are used for training models, which is an expensive process. Domain-adaptive approaches discussed in Section 1.2.2 to address the domain-shift issue, train the models, so that source domain data is supervised, and the target domain is unsupervised. The methods can be classified as, GAN-based methods, feature transfer methods, and methods with intermediate styles. Furthermore, in Chapter 4 of the thesis, the domain invariance has been addressed by using gradient reversal layer, to provide a direct alignment between source and target domains. However, the methods perform poorly in applications like in person re-identification which have greater challenges such as different camera configurations, different poses, and occlusions, which requires learning feature representations from the data, without explicit alignment.

Person re-identification (ReID) aims to solve the challenging task of identifying people across non-overlapping cameras. In many domain adaptation methods for person ReID, pseudo-labels are generated using clustering techniques, which suffer from noisy labels. In this chapter, we mitigate the noisy label problem through a novel contrastive feature learning method, performed in two steps. In the first step, weighted local temporal intra-region features are computed by integrating with region-specific augmentation. These intra-region features are combined spatially to generate inter-region global features. We performed our experimental evaluation on popular datasets including Market-1501, DukeMTMC-reID, MSMT17 and PersonX. The method proves its effectiveness by outperforming the best existing methods during evaluation. The content of this chapter

is currently under review and has been submitted to IEEE International Conference on Multimedia and Expo (ICME 2024)¹.

6.1 Introduction

Person re-identification (ReID) is used to identify persons across multiple non-overlapping cameras. This necessitates dealing with challenges such as lighting changes, pose changes, and occlusions (Wei et al., 2018). Another important challenge with person re-identification is the domain gap between the training data and the test data. To mitigate this issue, fully supervised methods require labeled data from both source and target domains during training. However, obtaining labeled data for the target domain is either impractical or often not possible in real-life scenarios. Hence, there is a need to develop unsupervised methods for learning ReID models using domain adaptation when annotations are not available for the target domain. The following section summarizes the related work in this field.

6.1.1 Related Work

To solve the problem of cross-domain ReID, unsupervised domain adaptive learning approaches have been proposed. This chapter focuses on the unsupervised domain adaptive (UDA) person ReID. One technique of achieving domain adaptation is by generation of domain invariant features to perform source and target domain feature alignment. This is achieved by minimizing the distance and aligning the source and target distribution (Sun et al., 2016; Kang et al., 2019) of source and target feature divergence, and reconstructing the source or target domain data (Ghifary et al., 2016).

Several style transfer approaches have been proposed to address the challenge of domain gap. SPGAN (Deng et al., 2018) proposes translation of image cues between source and target domain, that achieves self-similarity and domain dissimilarity between the source and the target domain images. CR-GAN (Chen et al., 2019b) renders the source domain images by using the knowledge and guidance of the target domain instances. PT-GAN (Wei et al., 2018) studies a transfer function between the source and target which can work well for a single camera, but not across many different cameras leading to different person styles. AT-Net (Liu et al., 2019a) uses an ensemble of GANs for handling

¹<https://2024.ieeicme.org/>

variations in camera-view, illumination, resolution by addressing the different factors affecting the transfer function from the source to target domain. PDA-Net (Li et al., 2019) uses pose and domain discriminators and generalizes the model to learn domain invariant features along with pose disentanglement. DG-Net++ (Zou et al., 2020) disentangles for both id and non-id related features, by introducing a shared appearance space, and separate structure spaces for target and source domains to handle variations such as pose, and position. Adversarial networks are hard to train and suffer from mode collapse, and training instability (Fu et al., 2019). Hence, they have been found to be inferior in performance compared to clustering or ensemble approaches.

The clustering approaches used for unsupervised cross-domain ReID have been shown to give better results compared with non-clustering algorithms. In these approaches, pre-training is performed on the source domain labeled dataset and clustering techniques like K -means (Lloyd, 1982) and DBSCAN (Ester et al., 1996) are used to generate clusters for the target domain, which are used to generate pseudo-labels (Saito et al., 2017). The pseudo-labels thus obtained, are then trained to learn target domain invariant features. SSG (Fu et al., 2019) uses pseudo-labels obtained from the DBSCAN clusters from body part features as well as whole body features. AD-Cluster (Zhai et al., 2020) generates augmented clusters to obtain a discriminative model. Other approaches such as SpCL (Ge et al., 2020b) improve on clustering by implementing an efficient hybrid memory. IDM (Dai et al., 2021) employs intermediate domains to bridge the gap between source and target domain. More recently Fourier augmentation (Isobe et al., 2021) with progressive domain adaptation has been used with clustering. Hybrid memory management helps to effectively manage this data. One of the main drawbacks of clustering is that, when the features are noisy, then the clusters and the pseudo-labels also tend to be noisy affecting training performance and person ReID accuracy. In addition, UDA ReID methods use unconstrained augmentation methods such as random cropping which contributes to noisy labels.

Ensemble approaches help mitigate the problem with noisy pseudo-labels. Ensembling methods such as self-ensembling (French et al., 2017) use a student-teacher network for assigning pseudo-labels for target data. Dual networks (Fu et al., 2019; Zhai et al., 2020)

or a combination of many popular networks (Bertocco et al., 2021) are also used in ensemble methods. The ensemble tries to obtain complementary features by training the networks simultaneously. Mutual Mean Teaching Network (MMT) (Ge et al., 2020a) leverages the idea of using both off-line hard and on-line soft pseudo-labels during training. AWB (Wang et al., 2022) proposed a dual network which improves on previous networks by obtaining complementary features using pre- and post-attention methods. While ensemble networks have provided better results, they required greater resources in both time and space for training. With the ensemble approach, the networks should be appropriately designed to eliminate the spread of the noisy labels to the networks. Furthermore, the networks must always complement each other. In addition, an ensemble of networks is more difficult to train and requires more training time when compared to non-ensemble techniques.

6.1.2 Contributions

The thesis tackles the problem of domain invariance, described in Section 1.5.3 which is caused by different distributions of training and testing data. While many methods in Section 1.5.3, have been promising in solving domain-shift problem, they have not been successful in more complicated person ReID application, where there are varying challenges like lighting, occlusions, varying backgrounds, camera angles, poses, where it would be beneficial to learn the feature representations that are informative across domains. Our proposed method has overcome the limitations of the existing methods by providing robust, less-noisy features, better domain generalization, and efficient memory management. Our contributions can be summarized as follows:

- (1) We combine region-specific augmentation temporally to obtain the intra-region features to learn variations within a region in a guided manner;
- (2) We design a self-supervised learning mechanism by combining intra-region features spatially to generate inter-region features;
- (3) We generate pseudo-labels for self-supervised ReID learning by clustering the inter-region features;
- (4) Our proposed method has been validated on various benchmarks, and outperformed state-of-the-art UDA person ReID methods.

6.2 Method

In the first section below, we provide a brief description of our framework, which is partly inspired by SpCL (Ge et al., 2020b). In the following sections, we describe our novel approach of region-specific feature learning. This involves two aspects: (1) generating features with region-specific augmentation, (2) a training strategy based on self-supervised contrastive learning that combines region-specific features both temporally (intra-region) and spatially (inter-region). The feature accumulation is performed in hybrid memory.

6.2.1 Overall Framework

As shown in Figure 6.1, which describes the overall framework of our method, labeled source and unlabeled target images are used. The images are divided into regions, augmented by region-specific augmentation, and used to train our network for person ReID learning. The region-specific features thus generated, are weight-averaged temporally by considering intra- and inter-region accumulation as shown in Figure 6.5. The intra-region and inter-region feature accumulations are shown as Steps 1 and 2, respectively. As shown in Figure 6.5, the hybrid memory is used to store accumulated features for an image as well as by region. The stored features for the target domain are then used for pseudo-label generation. For the source domain data, the cluster centroids are determined as an average of the class features and for the target domain data, a clustering algorithm is used to determine cluster centroids. For features not belonging to clusters, the corresponding images are treated as separate individual classes. Region-specific contrastive learning is used to learn the ReID model by incorporating source classes, and target domain pseudo-labels, and the network is trained using a region-specific contrastive loss function. Finally, hybrid memory is used to handle class level and instance level clusters of the target.

6.2.2 Region-Specific Feature Augmentation

The first step in the feature generation algorithm is to generate features by spatially constrained augmentation. We propose a novel region-specific augmentation method to generate region-specific features in which the region that the augmentation is confined to is specified as an input parameter at training time. Referring to Figure 6.2, we notionally

divide the entire image into three equal regions as named - head, torso and leg. Since, for most images of people, this is an intuitive approximation, we chose the approach of three equal regions. Figure 6.2 (2), (3), and (4) illustrate how an image of a person is augmented with spatial constraints during the training process so that region-specific features could be extracted. For instance, a specific region is augmented (for example, in the head region, Figure 6.2 (2)) and the resulting image is used to train the ReID network.

The contrastive loss function in training allows the network to distinguish images based on other regions within the human body and does not rely on features extracted from the head region. Similarly, when other regions of the image, such as the torso and legs as shown in Figure 6.2 (3) and (4) respectively are augmented, the contrastive loss function is adapted to train the network to learn representations based on regions of the image other than the one that is augmented. In this manner, an ensemble of networks is trained to learn region-specific feature representations of an image.

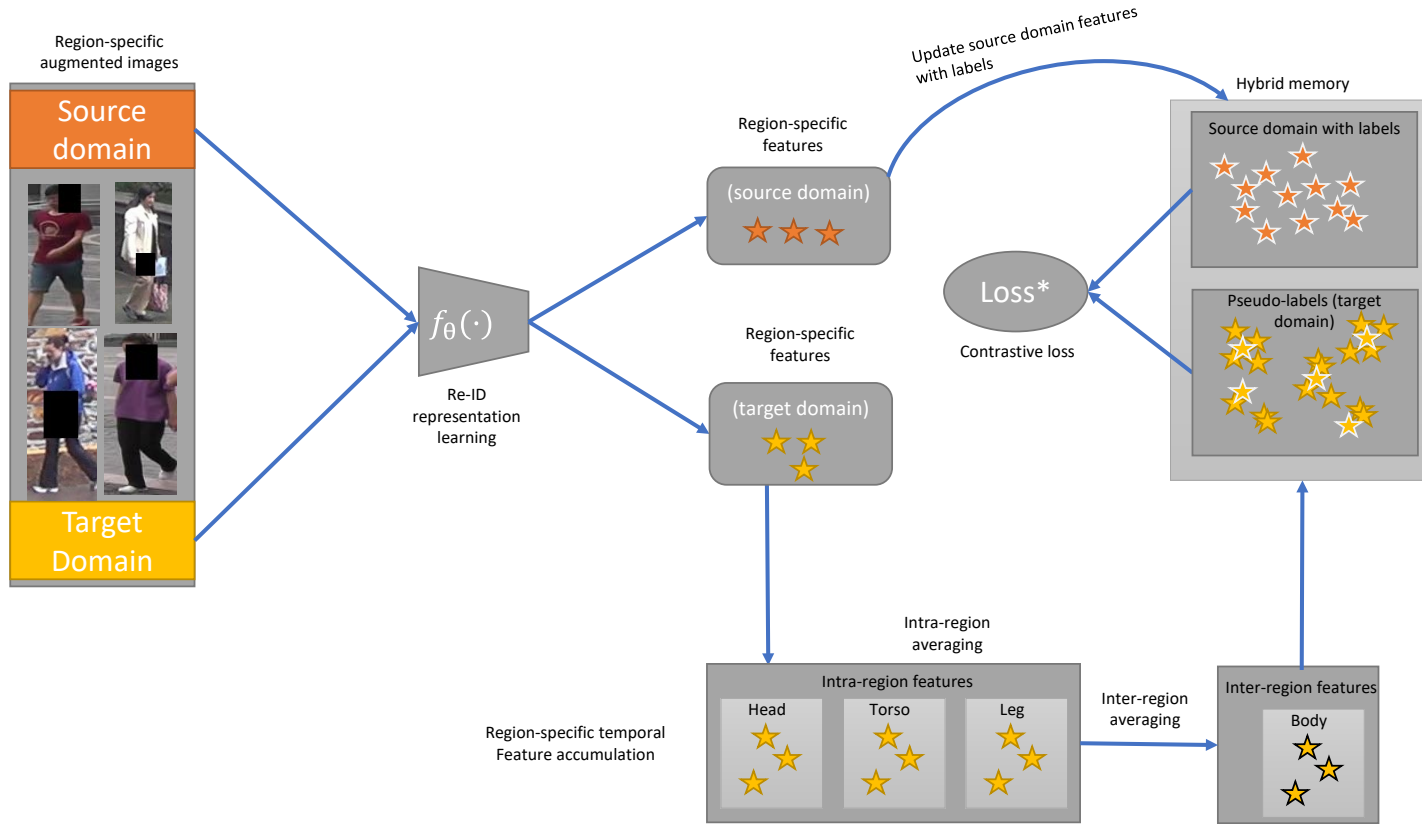


FIGURE 6.1: Architecture of our ReID learning framework which uses region-specific augmentation to train the network. The implementation of hybrid memory is illustrated to store region-specific features which are accumulated both temporally and spatially. The temporal accumulation creates intra-region features and spatial accumulation creates inter-region features. The cluster centers and pseudo-labels generated from ReID features learnt from region-specific augmentation are stored in hybrid memory.



FIGURE 6.2: An example image with the three identified regions: (1) the original RGB image, and region-specific augmentation is shown in (2) head region is augmented; (3) torso region is augmented and (4) leg region is augmented.

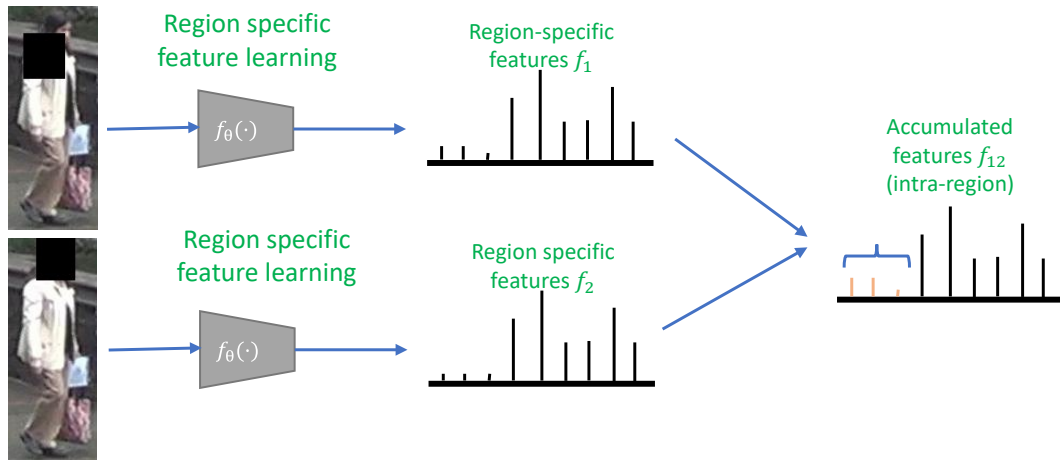


FIGURE 6.3: Region-specific features combined together to generate intra-region features. Two different augmentations of the head region are shown and the corresponding ReID features (f_1 , f_2) are combined to create intra-region features (f_{12}).

6.2.3 Feature Accumulation

A key intuition behind our approach is that temporal feature averaging or using historical features is equivalent to features generated by a network whose weights are averaged over

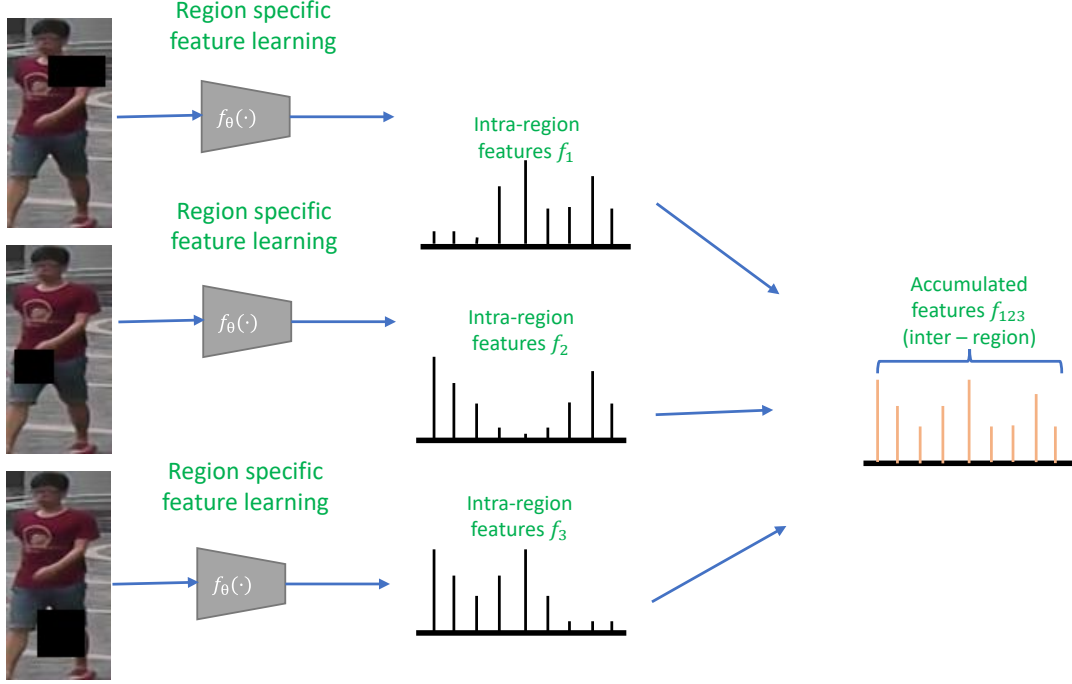


FIGURE 6.4: Region-specific features combined to produce an inter-region feature. Intra-region features which have been temporally accumulated in hybrid memory for head region (f_1), torso region (f_2) and leg region (f_3) are combined spatially to create an inter-region feature (f_{123}).

time (epochs). Both model and feature averaging can be viewed as ensemble learning methods where each ensemble member contributes an equal or weighted amount to the final prediction. We explain our approach to feature accumulation in the following two sections. The two feature accumulation methods are referred to as intra-region feature accumulation and inter-region feature accumulation.

6.2.3.1 Intra-Region Accumulation

For a given region, namely the head, torso, or leg, in each epoch, different augmentations are generated where the augmentation is confined to that single region. Figure 6.3 illustrates the situation where features are obtained from two separate augmentations of the head region, and combined temporally in a weighted manner to obtain an average feature at the end of each epoch. In the illustrated example, the learned feature representation f_1 is at the end of epoch 1 and f_2 is at the end of epoch 2, and both are combined to

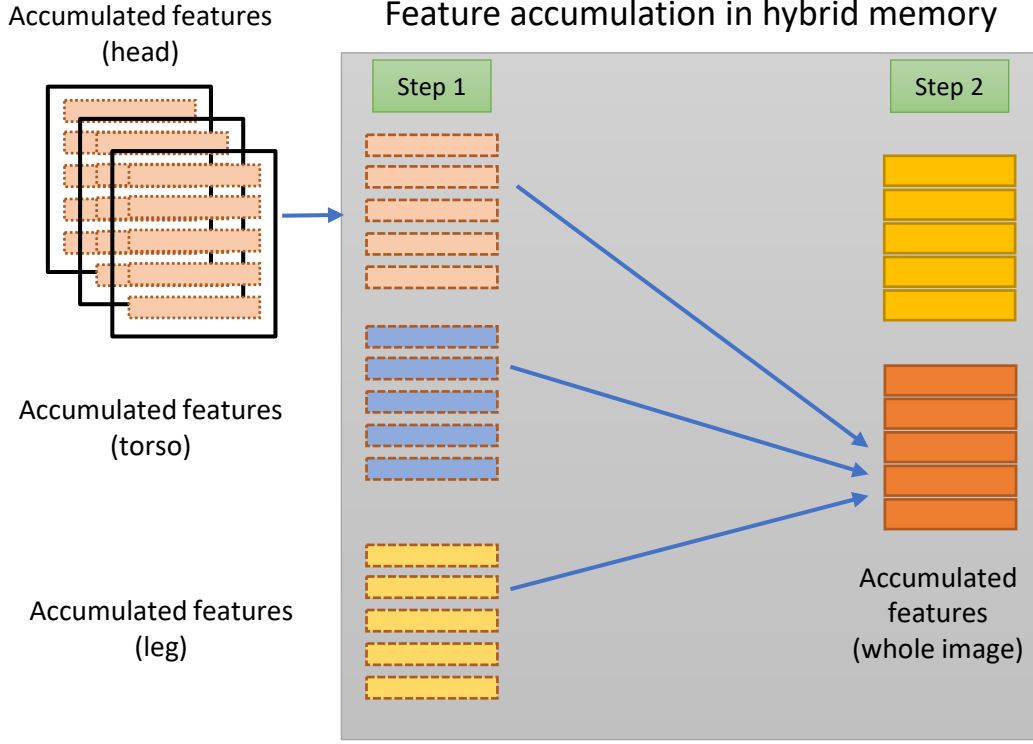


FIGURE 6.5: Outline of how region-specific features are accumulated in hybrid memory. In Step 1, intra-region features are accumulated during training, and in Step 2, the accumulated intra-region features are combined to form inter-region features.

obtain feature f_{12} using temporal weighting at the end of epoch 2. This is first step intra-region feature accumulation and the process is repeated for each region per epoch. This is furthermore illustrated in Figure 6.5 for all three regions of the image. The accumulation performed is described by the three equations that follow. $f_0(\theta)$ corresponds to the feature generated by the neural network where an augmentation is done within the head region. Similarly, $f_1(\theta)$ and $f_2(\theta)$ correspond to the torso and leg regions. Within a single epoch or a mini-batch, a specific region is chosen randomly. The generated feature is accumulated in hybrid memory separately for each such augmented region. For instance, the accumulated feature for the head is in f_0^{his} in Eq. (6.1), which represents the accumulated (averaged) temporal features of the two derived features, obtained after applying different augmentations as shown in Figure 6.3. In a similar manner, equations Eq. (6.2), and Eq. (6.3) correspond to the torso and the leg regions of the image:

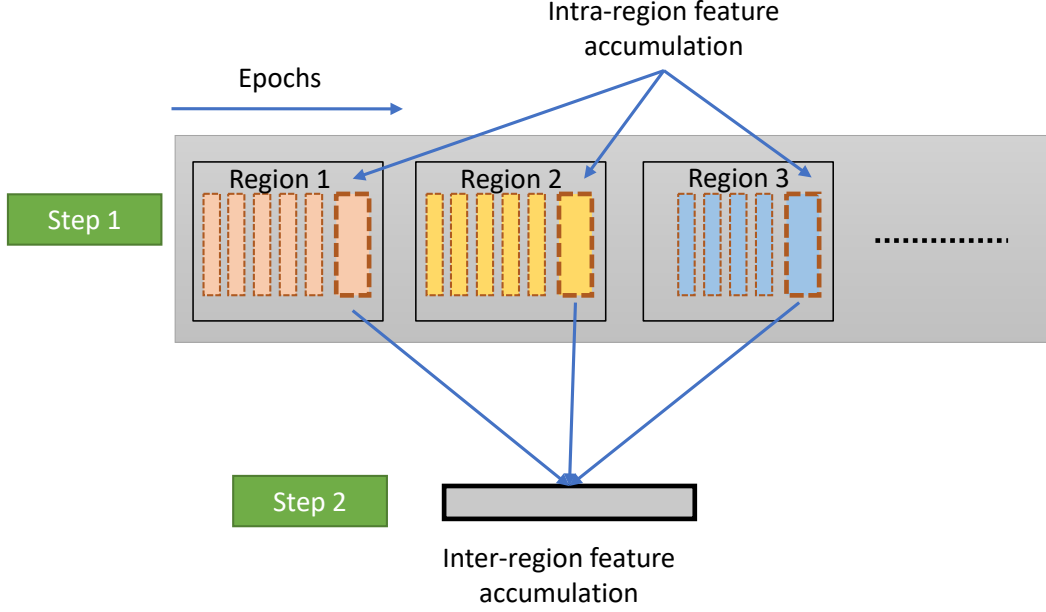


FIGURE 6.6: Implementation of feature accumulation in one experimental setting at the epoch level during training. Intra-region features are accumulated within the epoch. Inter-region features are obtained at the end of the epoch by combining intra-region features.

$$f_0^{his} = \alpha * f_0(\theta) + (1 - \alpha) * f_0^{his}, \quad (6.1)$$

$$f_1^{his} = \alpha * f_1(\theta) + (1 - \alpha) * f_1^{his}, \quad (6.2)$$

$$f_2^{his} = \alpha * f_2(\theta) + (1 - \alpha) * f_2^{his}. \quad (6.3)$$

6.2.3.2 Inter-Region Accumulation

The second step of feature accumulation is to perform spatial feature averaging across regions. Figure 6.4 illustrates the situation where features are obtained from augmentations in different regions of the image and accumulated. These intra-region features are then combined to form an overall average feature for the image. This process of inter-region feature accumulation is illustrated using equation Eq. (6.4), and Figure 6.4. In our approach, we take the average of all the region-specific accumulated features at the end of an epoch to determine the feature vector for the entire image. As shown in Figure 6.4, the

features f_1 , f_2 and f_3 are intra-region features that are accumulated temporally for three different regions of the image and then combined spatially as the inter-region feature f_{123} .

As shown in Figure 6.6, the temporally weighted intra-regions features obtained for each epoch for each region, are averaged to obtain inter-region feature, which is used to generate the pseudo-labels.

$$f^{his} = \frac{f_0^{his} + f_1^{his} + f_2^{his}}{3}. \quad (6.4)$$

6.3 Experiments and Results

This section provides a comprehensive overview of all the experiments, datasets and a comparison of the results with state-of-the-art methods and discussion of the results. In all experiments, the number of regions was set to 3 (0, 1, and 2), and the image was horizontally divided into equal thirds. Additionally, the weight α chosen for feature accumulation equations Eq. (6.1), (6.2) and (6.3) was empirically determined to be 0.8. Finally, the region augmentation chosen for the experiments was a novel random cropping which is confined to a region of the image. In the Section 6.3.2.1, we explain in detail the region-specific augmentation algorithm which we have used in our implementation.

The two-step feature accumulation as described earlier, was used in the main experiment. In the first part of the experiment, we generated region-specific features by choosing a region randomly with uniform probability and applying spatially constrained cropping as the augmentation for the region. The region-specific features were generated by using the same randomly chosen region for all images within an epoch, and the ReID network was trained with such images. Region-specific features were further combined with feature accumulation by performing two-step feature accumulation (intra-region and inter-region). At the end of an epoch, the intra-region features for that region were accumulated in hybrid memory. This process is illustrated in Figure 6.6 which graphically shows the two steps of feature accumulation. As seen in the figure, the region number used was kept constant during the epoch, and the features generated for the target domain images at the end of that epoch were maintained in separate memory within the hybrid memory.

For the ablation experiments, we have reported several variations of the main experiment in Section 6.3.3. In particular, we have explored different ReID backbones, and region choosing strategies.

6.3.1 Datasets

For our experimental evaluation, we used the same datasets that are used in the literature, so that we could make a fair comparison. As in the literature, we used Cumulative Matching Curve to compute evaluation metrics (Zheng et al., 2016b), with which, we computed Rank-1, Rank-5, and Rank-10. This, along with Mean Average Precision (mAP) (Zheng et al., 2015), was used for evaluation and comparison. Sample images from the four datasets are shown in Figure 6.7. Below is a brief description of the datasets.

Market-1501 (Zheng et al., 2015) (Market): This dataset has been collected at a supermarket in Tsinghua University using six cameras. A total of six cameras were used, including 5 high-resolution cameras. Field-of-view overlap exists among different cameras. The dataset has 12,936 images for training and 19,732 images for testing. The dataset also comprises 1501 identities, of which 750 are used in training and the other 751 are used in testing. All identities are captured by at least two cameras for cross-camera matching and up to six cameras in total.

DukeMTMC-reID (Ristani et al., 2016) (Duke): This dataset is a subset of DukeMTMC (Duke Multi-tracking multi-camera) for image-based person ReID. It consists of high-resolution person ReID images captured using 8 cameras. The dataset has 16,522 images of 702 identities for training. The test set has a total of 2,228 query images and 17,661 gallery images, covering another 702 identities.

MSMT17 (Wei et al., 2018) (MSMT): This dataset is a multi-scene dataset captured using 15 cameras, which is challenging due to the lighting variations and outdoor/indoor setting. The dataset also covers an extended period of time: four days in a month, three hours in the morning, noon and afternoon. Thus, the dataset has significant lighting and pose variations. It currently contains the largest number of annotated identities and bounding boxes. It comprises 32,621 images in the training set and 93,820 images in the testing set.



FIGURE 6.7: Sample images from the four datasets we have used in our evaluation: (a) Market-1501 (Zheng et al., 2015) dataset; (b) DukeMTMC-reID (Ristani et al., 2016) dataset; (c) MSMT17 (Wei et al., 2018) dataset showing variations in lighting, scene, pose and occlusions; (d) PersonX (Sun and Zheng, 2019) dataset with some sample images in various background and pose and the 3 colour and scene variations.

1401 identities are used for training and 3060 identities are available for testing. To date, this is one of the most challenging datasets.

PersonX dataset (Sun and Zheng, 2019): This is a synthetic dataset with 1266 identities with six different backgrounds. There are 30,816 images in the test set and 9840 images in the training set. In terms of background, there are three color backgrounds and three scene backgrounds. In terms of identities, there are 1266 in total (547 females and 719 males) and for each identity, 36 viewpoints are presented.

6.3.2 Implementation and Results

Our implementation is illustrated in Figure 6.1. The labels were used for the source domain and the no labels were used for the target domain. The input images were resized to be of dimensions $256 \times 128 \times 3$.

As mentioned earlier, clustering was used to generate pseudo-labels. We used ResNet50-IBN-a (Pan et al., 2018) as our backbone network pre-trained on ImageNet (Deng et al., 2009) with hybrid loss. Augmentation techniques such as random flipping, and our novel implementation of region-specific augmentation, were used. At the beginning of each epoch, DBSCAN with k-reciprocal nearest distance with Jaccard distance (Zhong et al., 2017) for clustering was used to generate pseudo-labels. The maximum distance between the points was set to 4. The model was trained with Adam Optimizer with a weight decay of 0.0005, and the initial learning rate was set to 0.00035. The temperature was empirically set to 0.04 and momentum to 0.3, and the model was trained for 40 epochs. No other pre-training or re-ranking methods were used.

6.3.2.1 Region-Specific Augmentation Algorithms

The region-specific augmentation algorithm uses random cropping as the augmentation method. The standard random cropping augmentation randomly crops a section of the image at training time. In the standard implementation, this crop can be of any size and shape, and can occur anywhere in the image in an unconstrained way. In contrast, our guided approach confines the augmentation to a specific region within the image.

The method to generate region-specific features is described through two algorithms, namely, *RegionAugment* and *ConfineAugmentation*. The algorithm *RegionAugment*, provides the overall procedure for obtaining a region-specific augmentation for an image, where image cropping is used as an augmentation method. This algorithm is adapted to generate suitable crops confined to specific regions with an image. The overall approach is to first generate a random crop, which can span any part of the image. This crop is then adjusted to be confined within a given selected region by the *ConfineAugmentation* algorithm. The detailed procedure for *RegionAugment* is shown in Algorithm 3.

The second algorithm, *ConfineAugmentation*, illustrates the implementation of confining an augmentation, cropping in this instance, to a given region, as shown in Figure 6.2. This algorithm adjusts the cropped region, by clamping the coordinates along the length of the image, so that the minimum and maximum extents are within the selected region.

The region is specified as an input parameter *region*, which, in this instance, assigns values in the range $[0, 3)$, which partitions the image vertically into equal portions. The detailed procedure for *ConfineAugmentation* is shown in Algorithm 4.

Algorithm 3 RegionAugment

Input: *image*

Output: *cropped_image*

```

1: while TRUE do
2:    $x_1, x_2, y_1, y_2 = \text{random\_crop}(\text{image})$ 

   // Try till a valid augmentation is obtained.
3:   if (not valid augmentation) then
4:     continue
5:   end if
   // Determine region
6:    $\text{part} = \text{determine\_region}()$ 

   // Confine augmentation within a region.
7:    $x_{p1}, x_{p2}, y_{p1}, y_{p2} = \text{ConfineAugmentation}(\text{imgw}, \text{imgw}, x_1, x_2, y_1, y_2, \text{part})$ 

   // Try till a valid crop within part is obtained.
8:   if (not valid augmentation in the region) then
9:     continue
10:  end if
11:   $\text{cropped\_image} = \text{image}[x_{p1}, x_{p2}, y_{p1}, y_{p2}]$ 

12:  break

13: end while
14: return cropped_image

```

Algorithm 4 ConfineAugmentation

Input: $w_{img}, h_{img}, x_1, x_2, y_1, y_2, region$
Output: $x_{p1}, x_{p2}, y_{p1}, y_{p2}$

```

1: if ( $region == 0$ ) then
2:    $y_{min} = 0$ 
3:    $y_{max} = h_{img}/3$ 
4: else if ( $region == 1$ ) then
5:    $y_{min} = h_{img}/3$ 
6:    $y_{max} = 2 * h_{img}/3$ 
7: else
8:    $y_{min} = 2 * h_{img}/3$ 
9:    $y_{max} = h_{img}$ 
10: end if
11: if ( $y_1 < y_{min}$ ) then
12:    $y_1 = y_{min}$ 
13: else if ( $y_1 > y_{max}$ ) then
14:    $y_1 = y_{max}$ 
15: end if
16: if ( $x_2 < x_{min}$ ) then
17:    $x_2 = x_{min}$ 
18: else if ( $x_2 > x_{max}$ ) then
19:    $x_2 = x_{max}$ 
20: end if
21:  $x_{p1} = x_1, x_{p2} = x_2, y_{p1} = y_1, y_{p2} = y_2$ 
22: return  $x_{p1}, x_{p2}, y_{p1}, y_{p2}$ 

```

6.3.2.2 Experimental Results

The comparison of our method with state-of-the-art methods for person ReID, which use domain adaptation, was performed on both real and synthetic datasets, and the results are shown in Tables 6.1, 6.2, 6.3, 6.5, and 6.4. The state-of-the-art methods used for comparison included MMT (Ge et al., 2020a), AD-Cluster (Zhai et al., 2020), SpCL (Ge et al., 2020a), IDM (Dai et al., 2021), TDR (Isobe et al., 2021) and CCTSE (Bertocco et al., 2021). These methods use multiple networks, clustering, hybrid solution and ensemble methods respectively. As our method does not use an ensemble explicitly, to make a fair comparison, when a non-ensemble result of the state-of-the-art method is available, comparison has been made against it. Otherwise, the result of the ensemble method has been used for comparison. We have used the same scenarios for making comparisons as in the literature. These include Market to MSMT, Duke to MSMT, Market to Duke, Duke to

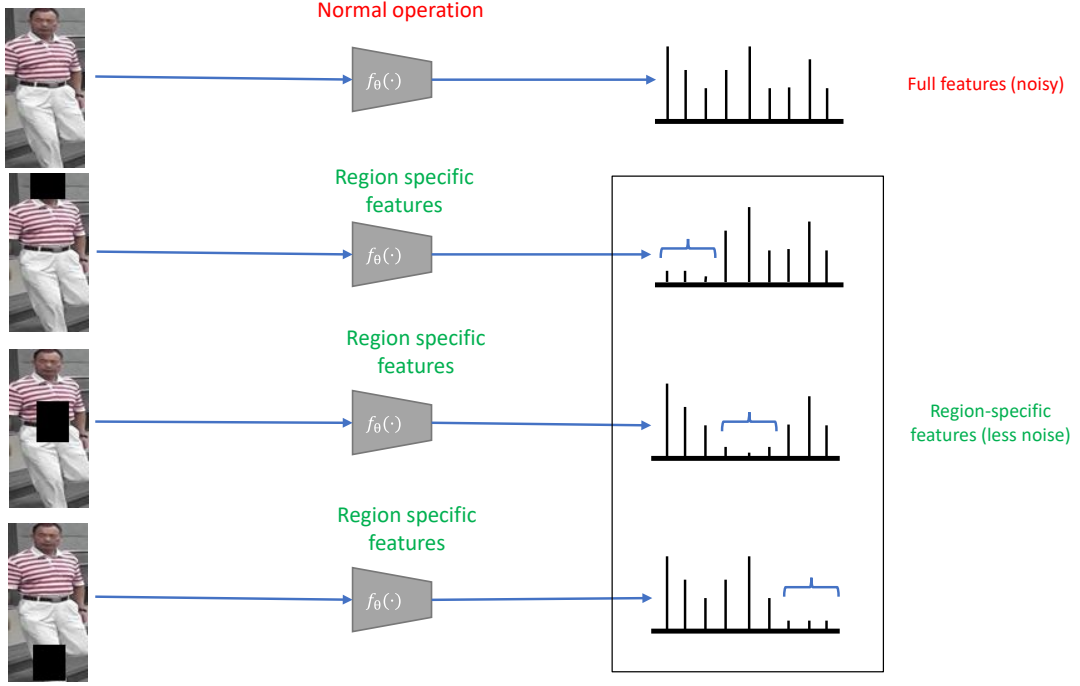


FIGURE 6.8: Region-specific features reduce noise as illustrated and hence result in robust clusters and pseudo-labels with less noise.

Market and PersonX to Market. In all the combinations we evaluated on, and the state-of-the-art methods we compared with, our method outperformed or performed close to state-of-the-art in Rank-1, Rank-5, Rank-10, and mAP.

The main disadvantage of the clustering step used in SpCL is error amplification due to noisy labels. In order to avoid training error amplification caused by noisy clusters, the self-paced learning strategy initialized the training process with the most reliable clusters which measured both compactness and distance between clusters. A limitation of this technique is the application of thresholds α and β for compactness and distance measures which need to be carefully chosen. Furthermore, the thresholds remain constant throughout the training phase and do not adapt to feature noise.

AD-Cluster uses density-based clustering generates the pseudo-labels, however, the clusters do not consider the source domain labels and hence lose valuable information. MMT requires two networks to train, adding to the training complexity when compared with our method. CCTSE requires an ensemble of networks to obtain satisfactory results. IDM utilizes an intermediate domain, which makes the network prone to noise by distorting the

TABLE 6.1: Results of our experiments with Duke as the source dataset and Market-1501 as the target dataset. We report mAP, RANK-1 (R1), RANK-5 (R5), RANK-10 (R10) results for our method, in comparison with state-of-the-art methods. The last row provides our implementation results. The best result is shown in bold, and the second-best result is underlined.

Duke To Market				
Methods	mAP	R1	R5	R10
MMT (Ge et al., 2020a)	71.2	87.7	94.9	96.9
AWB (Wang et al., 2022)	81.0	93.5	<u>97.4</u>	98.3
SpCL (Ge et al., 2020b)	76.7	90.3	96.2	97.7
AD-Cluster (Zhai et al., 2020)	68.3	86.7	94.4	96.5
CCTSE (Bertocco et al., 2021)	74.1	89.6	95.3	97.1
IDM (Dai et al., 2021)	<u>82.8</u>	<u>93.2</u>	97.5	<u>98.1</u>
TDR (Isobe et al., 2021)	82.2	93.6	-	-
Ours	83.5	93.0	97.5	98.3

information between the domains. TDR uses Fourier and spatial space, which introduces the complexity of handling multiple spaces. Furthermore, Fourier space may not be robust to variations in poses and scale.

In contrast, the novel features of our method provide considerable improvement when applied to a single network, and remove the dependency on training additional networks, Fourier spaces and intermediate domains. In our guided region-specific approach, the region-specific augmentation of the image as illustrated in Figure 6.8 results in features that have less noise. By masking portions of the image, the overall variation in the image is reduced, which leads to less noise in the learned features during training. This is illustrated graphically through the respective feature histograms as lesser variations with certain sections of the feature vector. This results in tighter and more compact clusters for the target domain images and hence less noisy labelling. From another perspective, by learning with the masked images, the model can generate discriminative features on

TABLE 6.2: Results of our experiments with Market-1501 as the source dataset and Duke as the target dataset. We report mAP, RANK-1 (R1), RANK-5 (R5), RANK-10 (R10) results for our method, in comparison with state-of-the-art methods. The last row provides our implementation results. The best result is shown in bold, and the second-best result is underlined.

Market To Duke				
Methods	mAP	R1	R5	R10
MMT (Ge et al., 2020a)	65.1	78.0	88.8	92.5
AWB (Wang et al., 2022)	<u>70.9</u>	83.8	92.3	94.0
SpCL (Ge et al., 2020b)	68.8	82.9	90.1	92.5
AD-Cluster (Zhai et al., 2020)	54.1	72.6	82.5	85.
CCTSE (Bertocco et al., 2021)	67.8	81.7	90.0	92.6
IDM (Dai et al., 2021)	70.5	<u>83.6</u>	<u>91.5</u>	93.7
TDR (Isobe et al., 2021)	69.4	82.7	-	-
Ours	71.3	83.8	91	<u>93.1</u>

TABLE 6.3: Results of our experiments with PersonX as the source dataset and Market-1501 as the target dataset. We report mAP, RANK-1 (R1), RANK-5 (R5), RANK-10 (R10) results for our method, in comparison with state-of-the-art methods. The last row provides our implementation results. The best result is shown in bold, and the second-best result is underlined.

PersonX to Market				
Methods	mAP	R1	R5	R10
MMT (Ge et al., 2020a)	71.0	86.5	94.8	97.0
SpCL (Ge et al., 2020b)	71.0	86.5	94.8	97.0
IDM (Dai et al., 2021)	81.3	92.0	97.4	98.2
TDR (Isobe et al., 2021)	<u>78.4</u>	<u>91.3</u>	-	-
Ours	78.1	92.9	<u>96.8</u>	<u>98.0</u>

TABLE 6.4: Results of our experiments with Market-1501 as the source dataset and MSMT as the target dataset. We report mAP, RANK-1 (R1), RANK-5 (R5), RANK-10 (R10) results for our method, in comparison with state-of-the-art methods. The last row provides our implementation results. The best result is shown in bold, and the second-best result is underlined.

Market to MSMT				
Methods	mAP	R1	R5	R10
MMT (Ge et al., 2020a)	24.0	50.1	63.5	69.3
AWB (Wang et al., 2022)	30.6	61.4	73.3	78.2
SpCL (Ge et al., 2020b)	31.0	58.1	69.6	74.1
CCTSE (Bertocco et al., 2021)	<u>33.2</u>	62.3	<u>74.1</u>	<u>78.5</u>
IDM (Dai et al., 2021)	33.5	61.3	73.9	78.4
TDR (Isobe et al., 2021)	32.9	<u>61.8</u>	-	-
Ours	33.5	64	75.7	79.7

the specific region which is masked. This reduces the noise in the pseudo-labels based on the masked regions, which is a benefit to the self-supervised learning step for the target domain.

6.3.3 Ablation Experiments

We conducted three ablation experiments by varying at least one aspect of the experimental setting described in the implementation section. In the first ablation experiment, the backbone of the network was changed from ResNet-50-IBN-a to ResNet50. The accuracy of results obtained by ResNet-50 is lower compared to those obtained by ResNet50-IBN-a, as shown in Tables 6.6 and 6.7, thus demonstrating that ResNet50-IBN-a has better performance compared to ResNet-50. Additionally, we wanted to evaluate if it was necessary to have region-specific feature generation performed across all three regions or just one region. Hence, we chose the torso covering region for guided feature generation, since we hypothesized that it has prominent features for comparison and greatest area.

TABLE 6.5: Results of our experiments with Duke as the source dataset and MSMT as the target dataset. We report mAP, RANK-1 (R1), RANK-5 (R5), RANK-10 (R10) results for our method, in comparison with state-of-the-art methods. The last row provides our implementation results. The best result is shown in bold, and the second-best result is underlined.

Duke to MSMT				
Methods	mAP	R1	R5	R10
MMT (Ge et al., 2020a)	23.3	50.1	63.9	69.8
AWB (Wang et al., 2022)	30.7	62.7	74.5	79.0
SpCL (Ge et al., 2020b)	26.5	53.1	65.8	70.5
CCTSE (Bertocco et al., 2021)	34.5	<u>63.9</u>	75.3	79.6
IDM (Dai et al., 2021)	35.4	63.6	<u>75.5</u>	<u>80.2</u>
TDR (Isobe et al., 2021)	32.7	62.7	-	-
Ours	<u>34.8</u>	66.4	77.7	81.5

TABLE 6.6: Results of mAP, RANK-1 (R1), RANK-5 (R5), and RANK-10 (R10) in the Market-1501 to Duke branch of the ablation experiment. The second row (ResNet-50) outlines the results when the backbone is ResNet-50. The third row (ResNet-50-Part) outlines the results when only one large region was used, and ResNet-50 was the backbone. The best results (ResNet50-IBN-a), shown in bold, are the first set of results with ResNet50-IBN-a as a backbone.

Market to Duke				
Methods	mAP	R1	R5	R10
ResNet50-IBN-a	71.3	83.8	91	93.1
ResNet-50	67.6	81.3	89.9	92.3
ResNet-50-Part	66.7	81.7	90.2	92.6

We then followed the same training procedure as the one involving ResNet-50. As shown in Tables 6.6 and 6.7 the accuracy of the result obtained by ResNet-50-Part, where guided feature learning was applied to the middle region, covering the torso only, is lower than

TABLE 6.7: Results of mAP, RANK-1 (R1), RANK-5 (R5), and RANK-10 (R10) in the Duke to Market-1501 branch of the ablation experiment. The second row (ResNet-50) outlines the results when the backbone is ResNet-50. The third row (ResNet-50-Part) outlines the results when only one large region was used, and ResNet-50 was the backbone. The best results (ResNet50-IBN-a), shown in bold, are the first set of results with ResNet50-IBN-a as a backbone.

Duke to Market				
Methods	mAP	R1	R5	R10
ResNet50-IBN-a	83.5	93.0	97.1	98.3
ResNet-50	77.0	90.7	96.6	97.8
ResNet-50-Part	74.7	89.5	95.4	96.8

the result obtained using all three region features. This implies that applying guided feature learning to all three regions, rather than one region, improved detection accuracy. It is hypothesized that this phenomenon would occur with other backbone networks.

In addition, we performed another set of ablation studies with MMT (Ge et al., 2020a) as the base framework. The results of the experiments are in Table 6.8. Referring to Figure 6.9, we conducted following variations of the approach as shown in Figure 6.6. This is shown in Figure 6.9 (b), Figure 6.9 (c) and Figure 6.9 (d). For easy comparison, Figure 6.9 (a) is same as Figure 6.6.

- ResNet-50-MMT: The baseline implementation of MMT.
- ResNet-50-Random-Region (Figure 6.9 (a)): This is explained in Section 6.2.
- ResNet-50-Random-Seq (Figure 6.9 (b)): in this method, the sequence of the regions is pre-determined. For each epoch, the intra-features are obtained from the pre-determined region in that same sequence order. In the Figure 6.9 (b), the regions Region 1, Region 2, and Region 3 are repeated.
- ResNet-50-Random-Epoch (Figure 6.9 (c)): in this method, within an epoch, different regions are randomly chosen for different mini-batches.
- ResNet-50-Region-One (Figure 6.9 (d)): in this method, the features are obtained from only one type of region. In our experiment, we have chosen the region covering the torso.

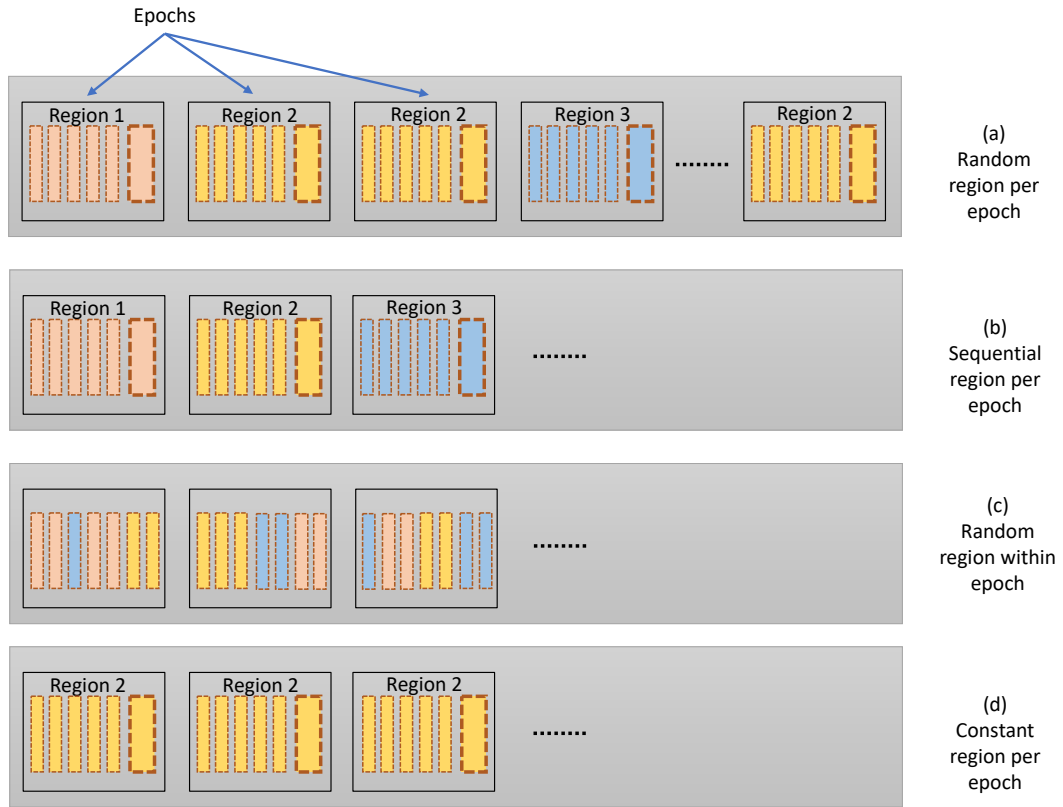


FIGURE 6.9: Ablation experiments: (a) A region is chosen randomly with uniform probability at the beginning of the epoch and used for region-specific augmentation for all images within that epoch; (b) A region is chosen in a sequential manner at the beginning of the epoch and applied for all images within the epoch to learn region-specific features; (c) A region is chosen randomly for each image within the epoch, thus creating different region-specific augmentations for individual images; (d) A constant region-specific augmentation is performed and in our case region 2 is chosen which is the torso.

The ablation results show that the approach shown in Figure 6.9 (a) provides better generalization than those shown in Figure 6.9 (b), Figure 6.9 (c) and Figure 6.9 (d). This implies that the features are generalized robustly in the method shown in Figure 6.9 (a) while the other approaches suffer from more noise (ResNet-50-Random-Epoch) or overfitting (ResNet-50-Random-Seq and ResNet-50-Region-One). This experiment also illustrates that our implementation, ResNet-50-Random-Region, provides better accuracy on other baseline implementations such as MMT with ResNet-50 as the backbone. Additionally, it also reinforces the generality and applicability of our method for different frameworks based on domain-adaptive methods for person ReID.

TABLE 6.8: Results of mAP, RANK-1 (R1), RANK-5 (R5), and RANK-10 (R10) in the Market-1501 to Duke branch performed on the MMT network. The first row has the results for two-stage feature accumulation method as part of the main experiment. The second row is baseline result for the MMT ReID network. Rows 3-5 are of the variations of our method as described in the Ablation Experiments section.

Methods	mAP	R1	R5	R10
ResNet-50-Random-Region (main experiment)	58.9	75.2	86.0	89.2
ResNet-50-MMT (baseline)	57.1	73.7	84.9	88.5
ResNet-50-Random-Seq	55.0	72.4	83.1	87.0
ResNet-50-Random-Epoch	56.6	73.9	85.0	88.1
ResNet-50-Region-One	56.8	73.7	84.3	88.6

6.4 Summary

Unsupervised domain adaptation techniques have been used for ReID, and clustering techniques have provided advantage of generating pseudo-labels for the target domain. While this is a sound technique, it has the disadvantage of noisy labels, which affect the training performance and overall accuracy. Our technique mitigates this by using region-specific random augmentation. The augmented images using our approach are used to generate weighted intra-features and inter-region features. The evaluation performed on datasets such as Market-1501, DukeMTMC-reID, MSMT17 and PersonX, have shown promising results and have provided improvement in the evaluation over state-of-the-art methods.

Part 3

Conclusions and Future Work

Conclusions and Future Work

7.1 Conclusions

Performance of deep neural network models rely on good quality annotated data, which is challenging and expensive to acquire, especially data which is representative of different domains and conditions. Furthermore, for better feature representation of the objects, pixel level annotation provides an advantage, by capturing intricate details and morphological features of the object. But this is not always available and impractical to create. This thesis proposes advanced learning methods to handle the following challenges: (1) imprecise localisation, (2) lack of adaptive generalization, (3) inadequate alignment of source and target domain data, (4) Weak and incomplete dataset annotations. Our methods try to solve these challenges by using both supervised and weakly supervised methods, as applicable to a variety of visual content analysis tasks.

In the first part of the thesis, we introduce supervised methods which address imprecise localisation and lack of generalization. In the later part of the thesis, we focus on weakly-supervised methods. We propose domain adaptive solutions for addressing domain gap as well as handling datasets with weak and limited annotations. We demonstrate the solution across diverse applications encompassing detection, re-identification and counting.

In Chapter 2, we propose methods to address inadequate generalization. Dropout methods are popular in deep neural networks to prevent over-fitting and learn a generalized model. However, they require tuning of hyper-parameters, structural changes, or ensemble of networks, which are all complex, time-consuming, and tedious. We propose two novel approaches to solve this problem. In the first approach, we determine the dropout rate automatically, by sampling from the distribution of the neuron activations. We extend this

approach, by adaptively determining the dropout rate by clustering the neuron activations and assigning a dropout rate for each cluster. Our methods outperform standard dropout and other state-of-the-art methods across all datasets and neural networks we evaluated on.

In Chapter 3, we introduce methods which show the advantage of precise localisation for objects and provide pixel-level masks for weakly annotated cell data. Mitosis cells in histopathology images are small, sparse, and have morphological variations which makes it difficult to detect without precise, localized masks. In this chapter, we introduce a method for mitosis cell detection in histopathology images. We use a combination of deep learned features with hand-crafted features to capture the features cells and outperform state-of-the-art methods in detection accuracy, thus proving that the feature representation based on pixel-level masks is better than the features obtained from bounding box localization for object detection. For weakly annotated datasets, which lack pixel-level annotation, we propose a novel method to also generate the masks automatically and use to for mitosis detection.

In Chapters 4 and 6 we study domain adaptive solutions to address the domain gap between the source and target datasets. In Chapter 4, the domain gap in the histopathology images occurs due to different staining, and tissue preparation, whilst in Chapter 6, the domain gap occurs due to different camera, lighting, and environmental conditions. In both these chapters, we tackle the challenge of classifying and detecting unlabeled objects in the target domain, where a fully annotated target domain dataset is not available.

In Chapter 4, we introduce an image and instance level dual-stage domain adaptive mitosis cell detection method, to handle the different stains of the histopathology images. We make use of the gradient reversal layer for domain adaptation and generate masks for localisation for improved domain adaptive mitosis detection. Additionally, our novel method automatically generates masks for a weakly annotated mitosis dataset, which was used as the target domain dataset. The method achieves better than state-of-the-art methods for mitosis detection and mask generation.

In Chapter 6, we provide a domain adaptive solution in the context of person re-identification application. In this chapter, we demonstrate a method of combining local and global features temporally and spatially using a novel region-specific augmentation. The features are used to generate pseudo-labels for the unlabeled target domain and the ReID network is trained in a self-supervised manner using contrastive loss. The method achieves state-of-the-art results for person re-identification on popular datasets.

In Chapter 5, we develop a novel method of cell counting using weakly annotated datasets with only dot annotations for cells. In our approach, we automatically generate and select good quality exemplars using cross-domain training. We exploit the cell structure similarity to achieve cross-domain knowledge transfer and apply it to generate exemplars automatically. Our method demonstrates that by using weakly-supervised training, where the training dataset is limited when compared to state-of-the-art methods, we can still obtain close to or better counting accuracy.

7.2 Future Work

This thesis focused on providing advanced deep learning solutions for computer vision for challenges listed in the introduction. We would like to propose some future work to take it forward.

Our supervised and weakly supervised methods for object detection, domain adaptation and counting can be extended and integrated into multi-modal solutions in various applications such as surveillance, digital pathology, agriculture. These multi-modal models can help mimic the actions of a human for decision making where information from all different modalities is considered and integrated before the final decision is made. For example, in digital pathology the data can be obtained not only from histopathology images but also from gene expressions, clinical information (Qiao et al., 2022). In yet another scenario, in surveillance, multi-modal information such as images, gait from videos, Infra-red can be integrated to improve the robustness of the applications (Zheng et al., 2021).

Our methods can be improved to make it more scalable and easily deployable by improving the architecture and framework, for practical purposes. For example, the methods

suggested for detection and counting, can be improved into a deployable single multi-task end-to-end solution (Liu et al., 2019b). Depending on the deployable environments and configurations, our architectures can be simplified to use light-weight CNN's like EfficientNet (Tan and Le, 2019). Furthermore, for the dropout method, the computational complexity of using clustering mechanism like K -means can be improved.

Transformers (Vaswani et al., 2017) have been successfully applied to Natural Language Processing (NLP) tasks. Given the growing practical importance of computer vision applications, a future research direction is to apply transformer-based architectures for addressing specific practical problems within computer vision including medical applications. As continually acquiring, and annotating new datasets becomes prohibitively expensive, we would like future research to focus on using transformers and attention-based architectures to be applied to unsupervised and weakly supervised settings. Advancements in this area will have a positive impact on many visual content analysis tasks such as person ReID, detection, and counting. Additionally, we note that, attention mechanism has been used successfully in supervised ReID learning (Li et al., 2018b). Hence, we propose new research direction focusing on adapting attention-based ReID in an unsupervised setting, include UDA based approaches.

Diffusion models (DM) (Sohl-Dickstein et al., 2015), which are built using a hierarchy of de-noising auto-encoders, have achieved impressive state-of-the-art results in image synthesis tasks. Diffusion model uses a Markov chain to progressively add noise to an image (forward diffusion process). Subsequently, through a process of reverse diffusion, which is a trainable process, the noise is removed to recover the original image. A deep neural network is used to learn the reverse of a Markovian diffusion process. However, training and inference using a neural network to model reverse diffusion has high computational complexity. Hence, more recently, Latent Diffusion Models (LDM) (Rombach et al., 2022), have been proposed, which have shown a lot promise in achieving optimal results in image synthesis tasks by applying diffusion in the latent representation rather than the pixel space. LDMs apply diffusion to the latent space of pre-trained auto-encoders, vastly reducing computation complexity while preserving high fidelity in image generation and achieving an optimal balance between computational complexity and image quality. On the back of the recent success of LDMs, they are now being studied in other

application domains such as image segmentation (Tomer Amit, 2021), object detection (Chen et al., 2023) and people counting (Ranasinghe et al., 2023). We hypothesise that a new research direction will emerge, adapting diffusion models for tackling the domain variance, a frequently occurring challenge in various advanced learning-based computer vision applications.

Bibliography

- Saad Ullah Akram, Juho Kannala, Lauri Eklund, and Janne Heikkilä. 2016. Cell segmentation proposal network for microscopy image analysis. In *Proceedings of the Deep Learning and Data Labeling for Medical Applications: First International Workshop, LABELS 2016, and Second International Workshop, DLMA 2016, held in conjunction with MICCAI 2016*, pages 21–29. Springer.
- Jon Almazan, Bojana Gajic, Naila Murray, and Diane Larlus. 2018. Re-id done right: Towards good practices for person re-identification. *CoRR*, abs/1801.05339.
- Giuseppe Amato, Luca Ciampi, Fabrizio Falchi, and Claudio Gennaro. 2019. Counting vehicles with deep learning in onboard uav imagery. In *IEEE Symposium on Computers and Communications*, pages 1–6. IEEE.
- Alexander Andreopoulos and John K Tsotsos. 2013. 50 years of object recognition: Directions forward. *Computer Vision and Image understanding*, 117(8):827–891.
- Carlos Arteta, Victor Lempitsky, J Alison Noble, and Andrew Zisserman. 2012. Learning to detect cells using non-overlapping extremal regions. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 348–356. Springer.
- Carlos Arteta, Victor Lempitsky, J Alison Noble, and Andrew Zisserman. 2014. Interactive object counting. In *Proceedings of the 13th European Conference on Computer Vision, Part III 13*, pages 504–518. Springer.
- Jimmy Ba and Brendan Frey. 2013. Adaptive dropout for training deep neural networks. In *Advances in Neural Information Processing Systems*, pages 3084–3092.
- Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495.

- K Sabeena Beevi, Madhu S Nair, and GR Bindu. 2019. Automatic mitosis detection in breast histopathology images using convolutional neural network based deep transfer learning. *Biocybernetics and Biomedical Engineering*, 39(1):214–223.
- Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175.
- Gabriel C Bertocco, Fernanda Andaló, and Anderson Rocha. 2021. Unsupervised and self-adaptative techniques for cross-domain person re-identification. *IEEE Transactions on Information Forensics and Security*, 16:4419–4434.
- Yuri Boykov, Olga Veksler, and Ramin Zabih. 2001. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239.
- Daniel Bug, Steffen Schneider, Anne Grote, Eva Oswald, Friedrich Feuerhake, Julia Schüller, and Dorit Merhof. 2017. Context-based normalization of histological stains using deep convolutional features. In *Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 3rd International Workshop, DLMIA, and 7th International Workshop, ML-CDS, held in conjunction with MICCAI*, pages 135–142. Springer.
- F.A. Hamprecht C. Sommer, L. Fiaschi and D.W. Gerlich. 2012. Learning-based mitotic cell detection in histopathological images. *International Conference on Pattern Recognition*, pages 2306–2309.
- Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. 2019. Exploring object relation in mean teacher for cross-domain detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11457–11466.
- Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslaine Gagnon. 2018. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353.

- Fabio Maria Cariucci, Lorenzo Porzi, Barbara Caputo, Elisa Ricci, and Samuel Rota Bulo. 2017. Autodial: Automatic domain alignment layers. In *IEEE International Conference on Computer Vision*, pages 5077–5085.
- Tianfeng Chai and Roland R Draxler. 2014. Root mean square error (rmse) or mean absolute error (mae). *Geoscientific Model Development Discussions*, 7(1):1525–1534.
- Antoni B Chan, Zhang-Sheng John Liang, and Nuno Vasconcelos. 2008. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7.
- Prithvijit Chattopadhyay, Ramakrishna Vedantam, Ramprasaath R Selvaraju, Dhruv Batra, and Devi Parikh. 2017. Counting everyday objects in everyday scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1135–1144.
- Dapeng Chen, Dan Xu, Hongsheng Li, Nicu Sebe, and Xiaogang Wang. 2018a. Group consistent similarity learning via deep crf for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8649–8658.
- Feng Chen, Michael P Pound, and Andrew P French. 2021. Learning to localise and count with incomplete dot-annotations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1612–1620.
- Guangyi Chen, Chunze Lin, Liangliang Ren, Jiwen Lu, and Jie Zhou. 2019a. Self-critical attention learning for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9637–9646.
- Hao Chen, Qi Dou, Xi Wang, Jing Qin, and Pheng Heng. 2016. Mitosis detection in breast cancer histology images via deep cascaded networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.
- Kai Chen, Donghai Guan, Weiwei Yuan, Bohan Li, Asad Masood Khattak, and Omar Alfandi. 2018b. A novel feature selection-based sequential ensemble learning method for class noise detection in high-dimensional data. In *Proceedings of the Advanced Data Mining and Applications: 14th International Conference*, pages 55–65. Springer.

- Ke Chen, Chen Change Loy, Shaogang Gong, and Tony Xiang. 2012. Feature mining for localised crowd counting. In *British Machine Vision Conference*, volume 1, page 3.
- Peibin Chen, Tao Ma, Xu Qin, Weidi Xu, and Shuchang Zhou. 2020. Data-efficient semi-supervised learning by reliable edge mining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9192–9201.
- Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. 2023. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19830–19843.
- Ting Chen and Christophe Chef d’Hotel. 2014. Deep learning based automatic immune cell detection for immunohistochemistry images. In *International Workshop on Machine Learning in Medical Imaging*, pages 17–24. Springer.
- Yanbei Chen, Massimiliano Mancini, Xiatian Zhu, and Zeynep Akata. 2022. Semi-supervised and unsupervised deep visual learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Yanbei Chen, Xiatian Zhu, and Shaogang Gong. 2019b. Instance-guided context rendering for cross-domain person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 232–242.
- Ying-Cong Chen, Xiatian Zhu, Wei-Shi Zheng, and Jian-Huang Lai. 2017. Person re-identification by camera correlation aware feature augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(2):392–408.
- Yuanyuan Chen and Zhang Yi. 2021. Adaptive sparse dropout: Learning the certainty and uncertainty in deep neural networks. *Neurocomputing*, 450:354–361.
- Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. 2018c. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3339–3348.
- Hyunjoo Cho, Sungbin Lim, Gunho Choi, and Hyunseok Min. 2017. Neural stain-style transfer learning using gan for histopathological images. *CoRR*, abs/1710.08543.

- Heung-Kook Choi, Torsten Jarkrans, Ewert Bengtsson, Janos Vasko, Kenneth Wester, Per-Uno Malmström, and Christer Busch. 1997. Image analysis based grading of bladder carcinoma. comparison of object, texture and graph based methods and their reproducibility. *Analytical Cellular Pathology*, 15(1):1–18.
- Hisham Cholakkal, Guolei Sun, Fahad Shahbaz Khan, and Ling Shao. 2019. Object counting and instance segmentation with image-level supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12397–12405.
- Francois Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Luca Ciampi, Giuseppe Amato, Fabrizio Falchi, Claudio Gennaro, and Fausto Rabitti. 2018. Counting vehicles with cameras. In *Symposium on Advanced Database System*.
- Krzysztof J Cios, Roman W Swiniarski, Witold Pedrycz, Lukasz A Kurgan, Krzysztof J Cios, Roman W Swiniarski, Witold Pedrycz, and Lukasz A Kurgan. 2007. Unsupervised learning: association rules. *Data Mining: A Knowledge Discovery Approach*, pages 289–306.
- D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber. 2013. Mitosis detection in breast cancer histology images with deep neural networks. *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 411–418.
- Yongxing Dai, Jun Liu, Yifan Sun, Zekun Tong, Chi Zhang, and Ling-Yu Duan. 2021. Idm: An intermediate domain module for domain adaptive person re-id. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11864–11874.
- Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *IEEE computer society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893. IEEE.
- J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

- Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. 2018. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 994–1003.
- Emily Denton, Sam Gross, and Rob Fergus. 2016. Semi-supervised learning with context-conditional generative adversarial networks. *CoRR*, abs/1611.06430.
- Terrance DeVries and Graham W Taylor. 2017. Improved regularization of convolutional neural networks with cutout. *CoRR*, abs/1708.04552.
- Nameirakpam Dhanachandra, Khumanthem Manglem, and Yambem Jina Chanu. 2015. Image segmentation using k-means clustering algorithm and subtractive clustering algorithm. *Procedia Computer Science*, 54:764–771.
- Najmeddine Dhieb, Hakim Ghazzai, Hichem Besbes, and Yehia Massoud. 2019. An automated blood cells counting and classification framework using mask r-cnn deep learning model. In *31st International Conference on Microelectronics (ICM)*, pages 300–303. IEEE.
- Lee R Dice. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.
- Veena Dodballapur, Rajanish Calisa, Yang Song, and Weidong Cai. 2020a. Automatic dropout for deep neural networks. In *Neural Information Processing: 27th International Conference, ICONIP 2020, Proceedings, Part III* 27, pages 185–196. Springer.
- Veena Dodballapur, Yang Song, Heng Huang, Mei Chen, Wojciech Chrzanowski, and Weidong Cai. 2019. Mask-driven mitosis detection in histopathology images. In *2019 IEEE 16th International Symposium on Biomedical Imaging*, pages 1855–1859. IEEE.
- Veena Dodballapur, Yang Song, Heng Huang, Mei Chen, Wojciech Chrzanowski, and Weidong Cai. 2020b. Dual-stage domain adaptive mitosis detection for histopathology images. In *Digital Image Computing: Techniques and Applications*, pages 1–7.

- Kunio Doi. 2007. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Computerized Medical Imaging and graphics*, 31(4-5):198–211.
- W Dong-DongChen and ZH WeiGao. 2018. Tri-net for semi-supervised deep learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 2014–2020.
- Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Knowledge Discovery and Data Mining*, volume 96, pages 226–231.
- Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338.
- J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray. 2015. Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012. *Int. J. of Cancer*, 136(5):E359–E386.
- David A Forsyth and Jean Ponce. 2002. *Computer vision: a modern approach*. prentice hall professional technical reference.
- Geoffrey French, Michal Mackiewicz, and Mark Fisher. 2017. Self-ensembling for visual domain adaptation. *CoRR*, abs/1706.05208.
- Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang. 2019. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *Proceedings of the IEEE/CVF international Conference on Computer Vision*, pages 6112–6121.
- Yarin Gal and Zoubin Ghahramani. 2015. Bayesian convolutional neural networks with bernoulli approximate variational inference. *CoRR*, abs/1506.02158.
- Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059. PMLR.

- Yaroslav Ganin and Victor Lempitsky. 2014. Unsupervised domain adaptation by back-propagation. In *Proceedings of the 32nd International Conference on Machine Learning ICML*, volume 37, pages 1180–1189.
- Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030.
- Yixiao Ge, Dapeng Chen, and Hongsheng Li. 2020a. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *International Conference on Learning Representations*.
- Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, et al. 2020b. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *Advances in Neural Information Processing Systems*, 33:11309–11321.
- Yonatan Geifman and Ran El-Yaniv. 2017. Deep active learning over the long tail. *CoRR*, abs/1711.00941.
- Niloofer Gheissari, Thomas B Sebastian, and Richard Hartley. 2006. Person reidentification using spatiotemporal appearance. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1528–1535. IEEE.
- Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. 2018. Dropblock: A regularization method for convolutional networks. *Advances in Neural Information Processing Systems*, 31.
- Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. 2016. Deep reconstruction-classification networks for unsupervised domain adaptation. In *Proceedings of the 14th European Conference, Part IV 14*, pages 597–613. Springer.
- Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448.
- Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587.

- Aidan N Gomez, Ivan Zhang, Siddhartha Rao Kamalakara, Divyam Madaan, Kevin Swersky, Yarin Gal, and Geoffrey E Hinton. 2019. Learning sparse networks using targeted dropout. *CoRR*, abs/1905.13678.
- Ian Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. 2013. Maxout networks. In *International Conference on Machine Learning*, pages 1319–1327. PMLR.
- Yue Guo, Oleh Krupa, Jason Stein, Guorong Wu, and Ashok Krishnamurthy. 2021. Sautnet: A unified network for cell counting in 2d and 3d microscopy images. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(4):1920–1932.
- Azam Hamidinekoo and Reyer Zwiggelaar. 2017. Stain colour normalisation to improve mitosis detection on breast histology images. In *Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 3rd International Workshop, DLMIA, and 7th International Workshop, ML-CDS, Held in Conjunction with MICCAI 2017*, pages 213–221. Springer.
- K. He, G. Gkioxari, P. Dollár, and R. Girshick. 2017. Mask R-CNN. *IEEE International Conference on Computer Vision*, pages 2980–2988.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1904–1916.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.
- Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580.
- H.Irshad. 2013. Automated mitosis detection in histopathology using morphological and multi-channel statistics features. *J.Pathol. Informat.*, 4:10–15.

- Meng-Ru Hsieh, Yen-Liang Lin, and Winston H Hsu. 2017. Drone-based object counting by spatially regularized regional proposal network. In *Proceedings of the IEEE international Conference on Computer Vision*, pages 4145–4153.
- Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. 2016. Deep networks with stochastic depth. In *European Conference on Computer Vision*, pages 646–661. Springer.
- Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2018. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5001–5009.
- Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456. pmlr.
- Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. 2019. Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5070–5079.
- Takashi Isobe, Dong Li, Lu Tian, Weihua Chen, Yi Shan, and Shengjin Wang. 2021. Towards discriminative representation learning for unsupervised person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8526–8536.
- Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2020. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2.
- Ni Jiang and Feihong Yu. 2020. Multi-column network for cell counting. *OSA Continuum*, 3(7):1834–1846.
- Chanho Jung and Changick Kim. 2010. Segmenting clustered nuclei using h-minima transform-based marker extraction and contour parameterization. *IEEE Transactions on Biomedical Engineering*, 57(10):2600–2604.

- Chanho Jung, Changick Kim, Seoung Wan Chae, and Sukjoong Oh. 2010. Unsupervised segmentation of overlapped nuclei using bayesian classification. *IEEE Transactions on Biomedical Engineering*, 57(12):2825–2832.
- Philipp Kainz, Martin Urschler, Samuel Schuler, Paul Wohlhart, and Vincent Lepetit. 2015. You should use regression to detect cells. In *Medical Image Computing and Computer-Assisted Intervention: 18th International Conference, Proceedings, Part III 18*, pages 276–283. Springer.
- Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. 2019. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4893–4902.
- Michael Kass, Andrew Witkin, and Demetri Terzopoulos. 1988. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331.
- Yan Ke and Rahul Sukthankar. 2004. Pca-sift: A more distinctive representation for local image descriptors. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004*, volume 2, pages II–II. IEEE.
- Rohit Keshari, Richa Singh, and Mayank Vatsa. 2019. Guided dropout. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4065–4072.
- Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. 2019. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12456–12465.
- Durk P Kingma, Tim Salimans, and Max Welling. 2015. Variational dropout and the local reparameterization trick. In *Advances in Neural Information Processing Systems*, pages 2575–2583.
- Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026.

- Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. 2012. Large scale metric learning from equivalence constraints. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2288–2295.
- Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. 2019. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1920–1929.
- Sonal Kothari, Qaiser Chaudry, and May D Wang. 2009. Automated cell counting and cluster segmentation using concavity detection and ellipse fitting techniques. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 795–798. IEEE.
- Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. 2009. Cifar-10 and cifar-100 datasets. URL: <https://www.cs.toronto.edu/kriz/cifar.html>, 6.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105.
- Krishna Kumar Singh and Yong Jae Lee. 2017. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3524–3533.
- Alex Labach, Hojjat Salehinejad, and Shahrokh Valaee. 2019. Survey of dropout methods for deep neural networks. *CoRR*, abs/1904.13310.
- Maxime W Lafarge, Josien PW Pluim, Koen AJ Eppenhof, Pim Moeskops, and Mitko Veta. 2017. Domain-adversarial neural networks to address the appearance variability of histopathology images. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 83–91. Springer.
- Issam H Laradji, Negar Rostamzadeh, Pedro O Pinheiro, David Vazquez, and Mark Schmidt. 2018. Where are the blobs: Counting by localization with point supervision. In *Proceedings of the European Conference on Computer Vision*, pages 547–562.

- Sanghun Lee and Chulhee Lee. 2020. Revisiting spatial dropout for regularizing convolutional neural networks. *Multimedia Tools and Applications*, 79(45-46):34195–34207.
- Victor Lempitsky and Andrew Zisserman. 2010. Learning to count objects in images. *Advances in Neural Information Processing Systems*, 23.
- C. Li, X. Wang, W. Liu, and L. J. Latecki. 2018a. Deep mitosis: Mitosis detection via deep detection, verification and segmentation networks. *Med. Imag. Anal.*, 45:121–133.
- Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. 2017a. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5542–5550.
- Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. 2017b. Identity-aware textual-visual matching with latent co-attention. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1890–1899.
- Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. 2014. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159.
- Wei Li, Xiatian Zhu, and Shaogang Gong. 2018b. Harmonious attention network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2285–2294.
- Y. Li, E. Mercan, S. Knezevitch, J. G. Elmore, and L. G. Shapiro. 2018c. Efficient and accurate mitosis detection-a lightweight rcnn approach. *International Conference on Pattern Recognition Applications and Methods*, pages 69–77.
- Yangkun Li, Weizhi Ma, Chong Chen, Min Zhang, Yiqun Liu, Shaoping Ma, and Yuekui Yang. 2022. A survey on dropout methods and experimental verification in recommendation. *IEEE Transactions on Knowledge and Data Engineering*.
- Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. 2017c. Fully convolutional instance-aware semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2359–2367.

- Yu-Jhe Li, Ci-Siang Lin, Yan-Bo Lin, and Yu-Chiang Frank Wang. 2019. Cross/-dataset person re-identification via unsupervised pose disentanglement and adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7919–7929.
- Dingkang Liang, Xiwu Chen, Wei Xu, Yu Zhou, and Xiang Bai. 2022. Transcrowd: Weakly-supervised crowd counting with transformers. *Science China Information Sciences*, 65(6):160104.
- Shengcai Liao and Stan Z Li. 2015. Efficient psd constrained asymmetric metric learning for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3685–3693.
- Hui Lin, Xiaopeng Hong, and Yabin Wang. 2021. Object counting: You only need to look at one. *CoRR*, abs/2112.05993.
- Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proceedings of the 13th European Conference, Part V 13*, pages 740–755. Springer.
- Fujun Liu and Lin Yang. 2017. A novel cell detection method using deep convolutional neural network and maximum-weight independent set. *Deep Learning and Convolutional Neural Networks for Medical Image Computing: Precision Medicine, High Performance and Large-Scale Datasets*, pages 63–72.
- Jiawei Liu, Zheng-Jun Zha, Di Chen, Richang Hong, and Meng Wang. 2019a. Adaptive transfer network for cross-domain person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7202–7211.
- Nianfeng Liu, Haiqing Li, Man Zhang, Jing Liu, Zhenan Sun, and Tieniu Tan. 2016a. Accurate iris segmentation in non-cooperative environments using fully convolutional

- networks. In *2016 International Conference on Biometrics*, pages 1–8. IEEE.
- Shikun Liu, Edward Johns, and Andrew J Davison. 2019b. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1871–1880.
- Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016b. Ssd: Single shot multibox detector. In *Proceedings of the 14th European Conference, Part I 14*, pages 21–37. Springer.
- Weizhe Liu, Mathieu Salzmann, and Pascal Fua. 2019c. Context-aware crowd counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5099–5108.
- Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. 2021a. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):857–876.
- Xiaofeng Liu, Bo Hu, Xiongchang Liu, Jun Lu, Jane You, and Lingsheng Kong. 2021b. Energy-constrained self-training for unsupervised domain adaptation. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7515–7520. IEEE.
- Stuart Lloyd. 1982. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137.
- De Rong Loh, Wen Xin Yong, Jullian Yapeter, Karupppasamy Subburaj, and Rajesh Chandramohanadas. 2021. A deep learning approach to the screening of malaria infection: Automated and rapid cell counting, object detection and instance segmentation using mask r-cnn. *Computerized Medical Imaging and Graphics*, 88:101845.
- Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015a. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440.
- Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. 2015b. Learning transferable features with deep adaptation networks. *CoRR*, abs/1502.02791.

- G Lowe. 2004. Sift-the scale invariant feature transform. *Int. J.*, 2(91-110):2.
- Erika Lu, Weidi Xie, and Andrew Zisserman. 2019. Class-agnostic counting. In *Computer Vision—ACCV 2018: 14th Asian Conference on Computer Vision, Part III 14*, pages 669–684. Springer.
- Pauline Luc, Camille Couprie, Soumith Chintala, and Jakob Verbeek. 2016. Semantic segmentation using adversarial networks. *CoRR*, abs/1611.08408.
- Roux Ludovic, Racocceanu Daniel, Loménie Nicolas, Kulikova Maria, Irshad Humayun, Klossa Jacques, Capron Frédérique, Genestie Catherine, et al. 2013. Mitosis detection in breast cancer histological images an icpr 2012 contest. *Journal of Pathology Informatics*, 4(1):8.
- Yucen Luo, Jun Zhu, Mengxi Li, Yong Ren, and Bo Zhang. 2018. Smooth neighbors on teacher graphs for semi-supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8896–8905.
- Andy J Ma, Pong C Yuen, and Jiawei Li. 2013. Domain transfer support vector ranking for person re-identification without target camera label information. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3567–3574.
- Laurens Maaten, Minmin Chen, Stephen Tyree, and Kilian Weinberger. 2013. Learning with marginalized corrupted features. In *International Conference on Machine Learning*, pages 410–418.
- Marc Macenko, Marc Niethammer, James S Marron, David Borland, John T Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E Thomas. 2009. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1107–1110. IEEE.
- Mausumi Maitra, Rahul Kumar Gupta, and Manali Mukherjee. 2012. Detection and counting of red blood cells in blood cell images using hough transform. *International Journal of computer applications*, 53(16).
- Mark Marsden, Kevin McGuinness, Suzanne Little, Ciara E Keogh, and Noel E O’Connor. 2018. People, penguins and petri dishes: Adapting object counting models

- to new visual domains and object types without forgetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8070–8079.
- Niki Martinel, Gian Luca Foresti, and Christian Micheloni. 2019. Aggregating deep pyramidal representations for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0.
- Maciej A Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. 2023. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 89:102918.
- Yanda Meng, Hongrun Zhang, Yitian Zhao, Xiaoyun Yang, Xuesheng Qian, Xiaowei Huang, and Yalin Zheng. 2021. Spatial uncertainty-aware semi-supervised crowd counting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15549–15559.
- Poorya Mianjy, Raman Arora, and Rene Vidal. 2018. On the implicit bias of dropout. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, volume 80, pages 3537–3545.
- Pietro Morerio, Jacopo Cavazza, Riccardo Volpi, René Vidal, and Vittorio Murino. 2017. Curriculum dropout. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3544–3552.
- Fabrice Muhlenbach, Stéphane Lallich, and Djamel A Zighed. 2004. Identifying and handling mislabelled instances. *Journal of Intelligent Information Systems*, 22(1):89–109.
- Abhirup Mukherjee, Nicole A Repina, David V Schaffer, and Ravi S Kane. 2017. Optogenetic tools for cell biological applications. *Journal of Thoracic Disease*, 9(12):4867.
- T Nathan Mundhenk, Goran Konjevod, Wesam A Sakla, and Kofi Boakye. 2016. A large contextual dataset for classification, detection and counting of cars with deep learning. In *Proceedings of the 14th European Conference, Part III 14*, pages 785–800. Springer.

- Tanguy Naets, Maarten Huijsmans, Paul Smyth, Laurent Sorber, and Gaël de Lannoy. 2021. A mask r-cnn approach to counting bacterial colony forming units in pharmaceutical development. *CoRR*, abs/2103.05337.
- Gireen Naidu, Tranos Zuva, and Elias Mmbongeni Sibanda. 2023. A review of evaluation metrics in machine learning algorithms. In *Computer Science On-line Conference*, pages 15–25. Springer.
- Laurent Najman and Michel Schmitt. 1994. Watershed of a continuous function. *Signal Processing*, 38(1):99–112.
- Mona Nashaat, Aindrila Ghosh, James Miller, and Shaikh Quader. 2021. Semi-supervised ensemble learning for dealing with inaccurate and incomplete supervision. *ACM Transactions on Knowledge Discovery from Data*, 16(3):1–33.
- Thanh Nguyen, Chau Pham, Khoi Nguyen, and Minh Hoai. 2022. Few-shot object counting and detection. In *European Conference on Computer Vision*, pages 348–365. Springer.
- Richard Nock and Frank Nielsen. 2004. Statistical region merging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(11):1452–1458.
- Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. 2015. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1520–1528.
- Mehdi Noroozi, Ananth Vinjimoor, Paolo Favaro, and Hamed Pirsiavash. 2018. Boosting self-supervised learning via knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9359–9367.
- Andreas Opelt, Axel Pinz, Michael Fussenegger, and Peter Auer. 2006. Generic object recognition with boosting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(3):416–431.
- Nobuyuki Otsu. 1979. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66.

- Zsolt Palotai, Miklós Láng, András Sárkány, Zoltán Tóssér, Daniel Sonntag, Takumi Toyama, and András Lőrincz. 2014. Labelmovie: Semi-supervised machine annotation tool with quality assurance and crowd-sourcing options for videos. In *2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–4. IEEE.
- Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. 2018. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision*, pages 464–479.
- Constantine Papageorgiou and Tomaso Poggio. 2000. A trainable system for object detection. *International Journal of Computer Vision*, 38:15–33.
- Joseph Paul Cohen, Genevieve Boucher, Craig A Glastonbury, Henry Z Lo, and Yoshua Bengio. 2017. Count-ception: Counting by fully convolutional redundant counting. In *Proceedings of the IEEE International Conference on Computer Vision workshops*, pages 18–26.
- Agne Paulauskaite-Taraseviciene, Kristina Sutiene, Justas Valotka, Vidas Raudonis, and Tomas Iesmantas. 2019. Deep learning-based detection of overlapping cells. In *Proceedings of the 3rd International Conference on Advances in Artificial Intelligence*, pages 217–220.
- Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. 2021. Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11557–11568.
- Nils Plath, Marc Toussaint, and Shinichi Nakajima. 2009. Multi-class image segmentation using conditional random fields and global classification. In *Proceedings of the 26th annual International Conference on Machine Learning*, pages 817–824.
- David MW Powers. 2020. Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.
- Lutz Prechelt. 2002. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer.

- Dong-Hyun Lee Pseudo-Label. 2013. The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop: Challenges in Representation Learning*, pages 1–6.
- Yixuan Qiao, Lianhe Zhao, Chunlong Luo, Yufan Luo, Yang Wu, Shengtong Li, Dechao Bu, and Yi Zhao. 2022. Multi-modality artificial intelligence in digital pathology. *Briefings in Bioinformatics*, 23(6):bbac367.
- Reza Moradi Rad, Parvaneh Saeedi, Jason Au, and Jon Havelock. 2018. Blastomere cell counting and centroid localization in microscopic images of human embryo. In *IEEE 20th International Workshop on Multimedia Signal Processing*, pages 1–6. IEEE.
- Reza Moradi Rad, Parvaneh Saeedi, Jason Au, and Jon Havelock. 2019. Cell-net: Embryonic cell counting and centroid localization via residual incremental atrous pyramid and progressive upsampling convolution. *IEEE Access*, 7:81945–81955.
- Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl Dickstein. 2017. On the expressive power of deep neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2847–2854.
- Yasiru Ranasinghe, Nithin Gopalakrishnan Nair, Wele Gedara Chaminda Bandara, and Vishal M Patel. 2023. Diffuse-denoise-count: Accurate crowd-counting with diffusion models. *arXiv preprint arXiv:2303.12790*.
- Viresh Ranjan and Minh Hoai Nguyen. 2022. Exemplar free class agnostic counting. In *Proceedings of the Asian Conference on Computer Vision*, pages 3121–3137.
- Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. 2021. Learning to count everything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3394–3403.
- Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788.
- Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojian Chen, and Xin Wang. 2021. A survey of deep active learning. *ACM computing*

- surveys*, 54(9):1–40.
- S. Ren, K. He, R. Girshick, and J. Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28:91–99.
- Zeyu Ren, Shuihua Wang, and Yudong Zhang. 2023. Weakly supervised machine learning. *CAAI Transactions on Intelligence Technology*.
- Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*, pages 17–35. Springer.
- Adrian Lopez Rodriguez and Krystian Mikolajczyk. 2019. Domain adaptation for object detection via style consistency. In *30th British Machine Vision Conference*, page 232.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of Medical Image Computing and Computer-Assisted Intervention, Part III 18*, pages 234–241. Springer.
- L. Roux, D. Racocceanu, N. Loménie, M. Kulikova, H. Irshad, J. Klossa, F. Capron, C. Genestie, G. Le Naour, and M. N. Gurcan. 2013. Mitosis detection in breast cancer histological images, an ICPR 2012 contest. *J. Pathol. Informat.*, 4.
- Aruni RoyChowdhury, Prithvijit Chakrabarty, Ashish Singh, SouYoung Jin, Huaizu Jiang, Liangliang Cao, and Erik Learned-Miller. 2019. Automatic adaptation of object detectors to new domains using self-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 780–790.
- Arnout C Ruifrok, Dennis A Johnston, et al. 2001. Quantification of histochemical staining by color deconvolution. *Analytical and Quantitative Cytology and Histology*, 23(4):291–299.

- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252.
- Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. 2008. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77:157–173.
- M. Saha, C. Chakraborty, and D. Racoceanu. 2018. Efficient deep learning model for mitosis detection using breast histopathology images. *Computerized Med. Imaging and Graphics*, 64:29–40.
- Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. 2017. Asymmetric tri-training for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 2988–2997. PMLR.
- Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. 2019. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6956–6965.
- Massimo Salvi, U Rajendra Acharya, Filippo Molinari, and Kristen M Meiburger. 2021. The impact of pre-and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis. *Computers in Biology and Medicine*, 128:104129.
- Meriem Sebai, Xinggang Wang, and Tianjiang Wang. 2020. Maskmitosis: a deep learning framework for fully supervised, weakly supervised, and unsupervised mitosis detection in histopathology images. *Medical & Biological Engineering & Computing*, 58:1603–1623.
- M Tarek Shaban, Christoph Baur, Nassir Navab, and Shadi Albarqouni. 2019. Staingan: Stain style transfer for digital histological images. In *2019 IEEE International Symposium on Biomedical Imaging (ISBI 2019)*, pages 953–956. IEEE.
- Connor Shorten and Taghi M Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48.

- Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C Kot, and Gang Wang. 2018. Dual attention matching network for context-aware feature sequence based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5363–5372.
- Rebecca L Siegel, Kimberly D Miller, Hannah E Fuchs, Ahmedin Jemal, et al. 2021. Cancer statistics, 2021. *Ca Cancer J Clin*, 71(1):7–33.
- K. Simonyan and A. Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.
- Vishwanath A Sindagi, Poojan Oza, Rajeev Yasarla, and Vishal M Patel. 2020. Prior-based domain adaptive object detection for hazy and rainy conditions. In *Proceedings of European Conference on Computer Vision, Part XIV 16*, pages 763–780. Springer.
- Saurabh Singh, Derek Hoiem, and David Forsyth. 2016. Swapout: Learning an ensemble of deep architectures. In *Advances in Neural Information Processing Systems*, pages 28–36.
- Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR.
- Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. 2022. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*.
- Carlos Oscar Sánchez Sorzano, Javier Vargas, and A Pascual Montano. 2014. A survey of dimensionality reduction techniques. *CoRR*, abs/1403.2877.
- MAI Sotelo, Ignacio Parra, David Fernandez, and Eugenio Naranjo. 2006. Pedestrian detection using svm and multi-feature combination. In *2006 IEEE Intelligent Transportation Systems Conference*, pages 103–108. IEEE.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

- J-L Starck, Michael Elad, and David L Donoho. 2005. Image decomposition via the combination of sparse representations and a variational approach. *IEEE Transactions on Image Processing*, 14(10):1570–1582.
- Baochen Sun, Jiashi Feng, and Kate Saenko. 2016. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.
- Xiaoxiao Sun and Liang Zheng. 2019. Dissecting person re-identification from the viewpoint of viewpoint. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 608–617.
- Esti Suryani, Wiharto Wiharto, and Nizomjon Polvonov. 2015. Identification and counting white blood cells and red blood cells using image processing case study of leukemia. *CoRR*, abs/1511.04934.
- Richard Szeliski. 2022. *Computer vision: algorithms and applications*. Springer Nature.
- Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR.
- Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 30.
- Ashkan Tashk, Mohammad Sadegh Helfroush, Habibollah Danyali, and Mojgan Akbarzadeh. 2013. An automatic mitosis detection method for breast cancer histopathology slide images based on objective and pixel-wise textural features classification. In *the 5th Conference on Information and Knowledge Technology*, pages 406–410. IEEE.
- F. B. Tek. 2013. Mitosis detection using generic features and an ensemble of cascade adaboosts. *J. Pathol. Informat.*, 4(1):12.
- David Tellez, Maschenka Balkenhol, Irene Otte-Höller, Rob van de Loo, Rob Vogels, Peter Bult, Carla Wauters, Willem Vreuls, Suzanne Mol, Nico Karssemeijer, et al. 2018. Whole-slide mitosis detection in h&e breast histology using phh3 as a reference

- to train distilled stain-invariant convolutional networks. *IEEE Transactions on Medical Imaging*, 37(9):2126–2136.
- J-P Thiran and Benoit Macq. 1996. Morphological feature extraction for the classification of digital images of cancerous tissues. *IEEE Transactions on Biomedical Engineering*, 43(10):1011–1020.
- Yingjie Tian and Yuqi Zhang. 2022. A comprehensive survey on regularization strategies in machine learning. *Information Fusion*, 80:146–166.
- Tal Shaharabany Lior Wolf Tomer Amit, Eliya Nachmani. 2021. Segdiff: Image segmentation with diffusion probabilistic models. *CoRR*, abs/2112.00390.
- Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. 2015. Efficient object localization using convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 648–656.
- Lindsey A Torre, Freddie Bray, Rebecca L Siegel, Jacques Ferlay, Joannie Lortet-Tieulent, and Ahmedin Jemal. 2015. Global cancer statistics, 2012. *CA: A Cancer Journal for Clinicians*, 65(2):87–108.
- Toan Tran, Trung Pham, Gustavo Carneiro, Lyle Palmer, and Ian Reid. 2017. A bayesian data augmentation approach for learning deep models. *Advances in Neural Information Processing Systems*, 30.
- Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. 2013. Selective search for object recognition. *International Journal of Computer Vision*, 104:154–171.
- Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang. 2016. A siamese long short-term memory architecture for human re-identification. In *Proceedings of the European Conference on Computer Vision, Part VII 14*, pages 135–153. Springer.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

- Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. 2019. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pages 6438–6447. PMLR.
- M. Veta, J. P. W. Pluim, P. J. Van Diest, and M. A. Viergever. 2014. Breast cancer histopathology image analysis: A review. *IEEE Transactions on Biomedical Engineering*, 61(5):1400–1411.
- M. Veta, P. J. van Diest, S. M. Willems, H. Wang, A. Madabhushi, A. Cruz-Roa, A. B. L. Larsen F. Gonzalez, J. S. Vestergaard, A. B. Dahl, et al. 2015. Assessment of algorithms for mitosis detection in breast cancer histopathology images. *Med. Imag. Anal.*, 20(1):237–248.
- Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the International Conference on Machine learning*, pages 1096–1103.
- Paul Viola and Michael Jones. 2001. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I. IEEE.
- Devesh Walawalkar, Zhiqiang Shen, Zechun Liu, and Marios Savvides. 2020. Attentive cutmix: An enhanced data augmentation approach for deep learning based image classification. In *International Conference on Acoustics, Speech and Signal Processing*, pages 3642–3646.
- Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. 2013. Regularization of neural networks using dropconnect. In *International Conference on Machine Learning*, pages 1058–1066.
- Guotai Wang, Wenqi Li, Sébastien Ourselin, and Tom Vercauteren. 2018. Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes, Held in Conjunction with MICCAI*, pages 178–190. Springer.

- H. Wang, A. Cruz-Roa, A. Basavanthally, H. Gilmore, N. Shih, M. Feldman, J. Tomaszewski, F. Gonzalez, and A. Madabhushi. 2014a. Cascaded ensemble of convolutional neural networks and handcrafted features for mitosis detection. *SPIE Medical Imaging*, 9041:90410B–90410B.
- Jianfeng Wang, Rong Xiao, Yandong Guo, and Lei Zhang. 2019a. Learning to count objects with few exemplar annotations. *CoRR*, abs/1905.07898.
- Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. 2016a. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12):2591–2600.
- Shengjie Wang, Abdel-rahman Mohamed, Rich Caruana, Jeff Bilmes, Matthai Philipose, Matthew Richardson, Krzysztof Geras, Gregor Urban, and Ozlem Aslan. 2016b. Analysis of deep neural networks with extended data jacobian matrix. In *International Conference on Machine Learning*, pages 718–726.
- Shengjie Wang, Tianyi Zhou, and Jeff Bilmes. 2019b. Bias also matters: Bias attribution for deep neural network explanation. In *International Conference on Machine Learning*, pages 6659–6667.
- Shengjie Wang, Tianyi Zhou, and Jeff Bilmes. 2019c. Jumpout: Improved dropout for deep neural networks with relus. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 6668–6676.
- Sida Wang and Christopher Manning. 2013. Fast dropout training. In *International Conference on Machine Learning*, pages 118–126.
- Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. 2014b. Person re-identification by video ranking. In *Computer Vision–ECCV 2014: Proceedings of the 13th European Conference on Computer Vision, Part IV 13*, pages 688–703. Springer.
- Wenhao Wang, Fang Zhao, Shengcai Liao, and Ling Shao. 2022. Attentive waveblock: Complementarity-enhanced mutual networks for unsupervised domain adaptation in person re-identification and beyond. *IEEE Transactions on Image Processing*.

- Xiao Wang, Daisuke Kihara, Jiebo Luo, and Guo-Jun Qi. 2020. Enaet: A self-trained framework for semi-supervised and supervised learning with ensemble transformations. *IEEE Transactions on Image Processing*, 30:1639–1647.
- Zuhui Wang. 2023. Cross-domain microscopy cell counting by disentangled transfer learning. In *International Workshop on Trustworthy Machine Learning for Healthcare*, pages 93–105. Springer.
- Zuhui Wang and Zhaozheng Yin. 2021. Annotation-efficient cell counting. In *Proceedings of Medical Image Computing and Computer Assisted Intervention, Part VIII 24*, pages 405–414. Springer.
- J. D. Webster and R. W. Dunstan. 2014. Whole-slide imaging and automated image analysis: considerations and opportunities in the practice of pathology. *Veterinary Pathology*, 51(1):211–223.
- Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. 2018. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 79–88.
- Adam S Wilkins and Robin Holliday. 2009. The evolution of meiosis from mitosis. *Genetics*, 181(1):3–12.
- Haibing Wu and Xiaodong Gu. 2015a. Max-pooling dropout for regularization of convolutional neural networks. In *2015, Neural Information Processing: 22nd International Conference, ICONIP, Proceedings, Part I 22*, pages 46–54. Springer.
- Haibing Wu and Xiaodong Gu. 2015b. Towards dropout training for convolutional neural networks. *Neural Networks*, 71:1–10.
- Xiongwei Wu, Doyen Sahoo, and Steven CH Hoi. 2020. Recent advances in deep learning for object detection. *Neurocomputing*, 396:39–64.
- Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747.

- Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698.
- Weidi Xie, J Alison Noble, and Andrew Zisserman. 2018. Microscopy cell counting and detection with fully convolutional regression networks. *Computer methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 6(3):283–292.
- Jiaolong Xu, Sebastian Ramos, David Vázquez, and Antonio M López. 2014. Domain adaptation of deformable part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(12):2367–2380.
- Jingyi Xu, Hieu Le, Vu Nguyen, Viresh Ranjan, and Dimitris Samaras. 2023. Zero-shot object counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15548–15557.
- Rui Xu and Donald Wunsch. 2005. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678.
- Fan Yang, Ke Yan, Shijian Lu, Huizhu Jia, Xiaodong Xie, and Wen Gao. 2019. Attention driven person re-identification. *Pattern Recognition*, 86:143–155.
- Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. 2022. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*.
- Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. 2021. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2872–2893.
- Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. 2014. Deep metric learning for person re-identification. In *2014 22nd International Conference on Pattern Recognition*, pages 34–39. IEEE.
- Yading Yuan, Ming Chao, and Yeh-Chi Lo. 2017. Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance. *IEEE Transactions on Medical Imaging*, 36(9):1876–1886.

- Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032.
- Muhamad Munawar Yusro, Rozniza Ali, and Muhammad Suzuri Hitam. 2023. Blood cell detection and counting in microscopy images using overlapping object recognizing model. In *AIP Conference Proceedings*, volume 2484. AIP Publishing.
- Sergey Zagoruyko and Nikos Komodakis. 2016. Wide residual networks. *CoRR*, abs/1605.07146.
- Yunpeng Zhai, Shijian Lu, Qixiang Ye, Xuebo Shan, Jie Chen, Rongrong Ji, and Yonghong Tian. 2020. Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9021–9030.
- Chao Zhang, Changming Sun, Ran Su, and Tuan D Pham. 2012. Segmentation of clustered nuclei based on curvature weighting. In *Proceedings of the 27th Conference on Image and Vision Computing New Zealand*, pages 49–54.
- Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. 2015. Cross-scene crowd counting via deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 833–841.
- Dan Zhang, Jingjing Li, Lin Xiong, Lan Lin, Mao Ye, and Shangming Yang. 2019. Cycle-consistent domain adaptive faster rcnn. *IEEE Access*, 7:123903–123911.
- Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018*,.
- Xin Zhang, Yee-Hong Yang, Zhiguang Han, Hui Wang, and Chao Gao. 2013. Object class detection: A survey. *ACM Computing Surveys*, 46(1):1–53.
- Aihua Zheng, Zi Wang, Zihan Chen, Chenglong Li, and Jin Tang. 2021. Robust multi-modality person re-identification. In *Proceedings of the AAAI Conference on Artificial*

Intelligence, volume 35, pages 3529–3537.

Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. 2016a. Mars: A video benchmark for large-scale person re-identification. In *Proceedings of the 14th European Conference, Proceedings, Part VI 14*, pages 868–884. Springer.

Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1116–1124.

Liang Zheng, Yi Yang, and Alexander G Hauptmann. 2016b. Person re-identification: Past, present and future. *CoRR*, abs/1610.02984.

Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. 2017a. Person re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1367–1376.

Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. 2011. Person re-identification by probabilistic relative distance comparison. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 649–656. IEEE.

Yushan Zheng, Zhiguo Jiang, Haopeng Zhang, Fengying Xie, Jun Shi, and Chenghai Xue. 2019. Adaptive color deconvolution for histological wsi normalization. *Computer Methods and Programs in Biomedicine*, 170:107–120.

Zhedong Zheng, Liang Zheng, and Yi Yang. 2017b. A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 14(1):1–20.

Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. 2017. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1318–1327.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. 2020. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13001–13008.

- Zhi-Hua Zhou. 2018. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53.
- Pengfei Zhu, Jingqing Li, Bing Cao, and Qinghua Hu. 2023. Multi-task credible pseudo-label learning for semi-supervised crowd counting. *IEEE Transactions on Neural Networks and Learning Systems*.
- Jingwei Zhuo, Jun Zhu, and Bo Zhang. 2015. Adaptive dropout rates for learning with corrupted features. In *24th International Joint Conference on Artificial Intelligence*.
- Konrad Zolna, Devansh Arpit, Dendi Suhubdy, and Yoshua Bengio. 2017. Fraternal dropout. *CoRR*, abs/1711.00066.
- Yang Zou, Xiaodong Yang, Zhiding Yu, BVK Kumar, and Jan Kautz. 2020. Joint disentangling and adaptation for cross-domain person re-identification. In *European Conference on Computer Vision*, pages 87–104. Springer.
- Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. 2019. Confidence regularized self-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5982–5991.