Configurable convolutional neural networks : Applications to breast cancer explainable classification and display panel defect detection

Feng He

N°d'ordre NNT : 2023ISAL0022

THESE de DOCTORAT DE L'INSA LYON, membre de l'UNIVERSITE DE LYON

En cotutelle internationale avec
Harbin Institute of Technology

Ecole Doctorale N°ED160
Electronique, électrotechnique, automatique

Spécialité / discipline de doctorat :
Traitement du Signal et de l'Image

Soutenue publiquement le 29/03/2023, par :
Feng HE

Configurable Convolutional Neural Networks: Applications to Breast Cancer Explainable Classification and Display Panel Defect Detection

Devant le jury composé de :

MME VINCENT Nicole, Professeur à Laboratoire LIPADE	Rapporteuse
M. YANG Jie, Professeur à Shanghai Jiao Tong University	Rapporteur
M. DUPONT Florent, Professeur à Université Lyon 1	Examinateur
M. LIU Wanyu, Professeur à Shanghai University	Examinateur
M. ZHU Yuemin, Directeur de Recherche CNRS à INSA de Lyon	Directeur de thèse
M. LIU Zhengjun, Professeur à Harbin Institute of Technology	Co-directeur de thèse

INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

Département de la Formation par la Recherche et des Études Doctorales (**FEDORA**)

Bâtiment INSA direction, 1er étage
37, av. J. Capelle
69621 Villeurbanne Cédex
fedora@insa-lyon.fr

Référence : TH0958_HE

L'INSA Lyon a mis en place une procédure de contrôle systématique via un outil de détection de similitudes (logiciel Compilatio). Après le dépôt du manuscrit de thèse, celui-ci est analysé par l'outil. Pour tout taux de similarité supérieur à 10%, le manuscrit est vérifié par l'équipe de FEDORA. Il s'agit notamment d'exclure les auto-citations, à condition qu'elles soient correctement référencées avec citation expresse dans le manuscrit.

Par ce document, il est attesté que ce manuscrit, dans la forme communiquée par la personne doctorante à l'INSA Lyon, satisfait aux exigences de l'Etablissement concernant le taux maximal de similitude admissible.

Département FEDORA – INSA Lyon - Ecoles Doctorales

SIGLE	ECOLE DOCTORALE	NOM ET COORDONNEES DU RESPONSABLE
CHIMIE	**CHIMIE DE LYON** **https://www.edchimie-lyon.fr** Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage secretariat@edchimie-lyon.fr	**M. Stéphane DANIELE** C2P2-CPE LYON-UMR 5265 Bâtiment F308, BP 2077 43 Boulevard du 11 novembre 1918 69616 Villeurbanne **directeur@edchimie-lyon.fr**
E.E.A.	**ÉLECTRONIQUE, ÉLECTROTECHNIQUE, AUTOMATIQUE** **https://edeea.universite-lyon.fr** Sec. : Stéphanie CAUVIN Bâtiment Direction INSA Lyon Tél : 04.72.43.71.70 secretariat.edeea@insa-lyon.fr	**M. Philippe DELACHARTRE** INSA LYON Laboratoire CREATIS Bâtiment Blaise Pascal, 7 avenue Jean Capelle 69621 Villeurbanne CEDEX Tél : 04.72.43.88.63 **philippe.delachartre@insa-lyon.fr**
E2M2	**ÉVOLUTION, ÉCOSYSTÈME, MICROBIOLOGIE, MODÉLISATION** **http://e2m2.universite-lyon.fr** Sec. : Bénédicte LANZA Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 secretariat.e2m2@univ-lyon1.fr	**Mme Sandrine CHARLES** Université Claude Bernard Lyon 1 UFR Biosciences Bâtiment Mendel 43, boulevard du 11 Novembre 1918 69622 Villeurbanne CEDEX **sandrine.charles@univ-lyon1.fr**
EDISS	**INTERDISCIPLINAIRE SCIENCES-SANTÉ** **http://ediss.universite-lyon.fr** Sec. : Bénédicte LANZA Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 secretariat.ediss@univ-lyon1.fr	**Mme Sylvie RICARD-BLUM** Institut de Chimie et Biochimie Moléculaires et Supramoléculaires (ICBMS) - UMR 5246 CNRS - Université Lyon 1 Bâtiment Raulin - 2ème étage Nord 43 Boulevard du 11 novembre 1918 69622 Villeurbanne Cedex Tél : +33(0)4 72 44 82 32 **sylvie.ricard-blum@univ-lyon1.fr**
INFOMATHS	**INFORMATIQUE ET MATHÉMATIQUES** **http://edinfomaths.universite-lyon.fr** Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage Tél : 04.72.43.80.46 infomaths@univ-lyon1.fr	**M. Hamamache KHEDDOUCI** Université Claude Bernard Lyon 1 Bât. Nautibus 43, Boulevard du 11 novembre 1918 69 622 Villeurbanne Cedex France Tél : 04.72.44.83.69 **hamamache.kheddouci@univ-lyon1.fr**
Matériaux	**MATÉRIAUX DE LYON** **http://ed34.universite-lyon.fr** Sec. : Yann DE ORDENANA Tél : 04.72.18.62.44 yann.de-ordenana@ec-lyon.fr	**M. Stéphane BENAYOUN** Ecole Centrale de Lyon Laboratoire LTDS 36 avenue Guy de Collongue 69134 Ecully CEDEX Tél : 04.72.18.64.37 **stephane.benayoun@ec-lyon.fr**
MEGA	**MÉCANIQUE, ÉNERGÉTIQUE, GÉNIE CIVIL, ACOUSTIQUE** **http://edmega.universite-lyon.fr** Sec. : Stéphanie CAUVIN Tél : 04.72.43.71.70 Bâtiment Direction INSA Lyon mega@insa-lyon.fr	**M. Jocelyn BONJOUR** INSA Lyon Laboratoire CETHIL Bâtiment Sadi-Carnot 9, rue de la Physique 69621 Villeurbanne CEDEX **jocelyn.bonjour@insa-lyon.fr**
ScSo	**ScSo*** **https://edsciencessociales.universite-lyon.fr** Sec. : Mélina FAVETON INSA : J.Y. TOUSSAINT Tél : 04.78.69.77.79 melina.faveton@univ-lyon2.fr	**M. Bruno MILLY** Université Lumière Lyon 2 86 Rue Pasteur 69365 Lyon CEDEX 07 **bruno.milly@univ-lyon2.fr**

*ScSo : Histoire, Géographie, Aménagement, Urbanisme, Archéologie, Science politique, Sociologie, Anthropologie

Abstract

Current deep learning methods such as convolutional neural networks (CNNs) are often dedicated to a specific task and object; they are generally fixed in network architecture, which limits their generalizability and prevents them from addressing multiple scenarios with different objectives. To achieve both the explainable classification of breast cancer and the online defect detection of display panels, we propose a configurable convolutional neural network (ConfigNet) capable of being transformed into different configurations according to the tasks and objects in question. The ConfigNet presents two main functional configurations. The first is composed of a feature extraction module (FEM), a decision map generator (DMG) and a classifier; it is devoted to image explainable classification, for which we propose two DMG structures and a weighted average pooling (WAP) classifier for histopathological breast cancer images. The second is an encoder-decoder configuration devoted to object segmentation and localization. In this second configuration, we propose an efficiency-favored decoder and an element-wise feature fusion module (EFFM) guiding the skip connection between the encoder and decoder for online defect detection of display panels. In addition, we develop a spatial and channel attention-guided feature fusion module (SCAFFM) and a bottleneck-structured decoder for breast tumor segmentation. The FEM or encoder in these two configurations is constructed through transfer learning from existing CNNs having deep convolutional layers.

Keywords— Deep learning, Configurable convolutional neural network, Weakly-supervised learning, Explainable classification, XAI, Segmentation, Breast cancer, Histopathological images, Defect detection, Display panel

Résumé

Les méthodes actuelles d'apprentissage profond, telles que les réseaux de neurones convolutifs (CNNs), sont souvent dédiées à une tâche et à un objet spécifiques ; leur architecture de réseau est généralement fixe, ce qui limite leur généralisabilité et les empêche d'aborder de multiples scénarios avec des objectifs différents. Pour réaliser à la fois la classification explicable du cancer du sein et la détection en ligne des défauts des panneaux d'affichage, nous proposons un réseau de neurones convolutif configurable (ConfigNet) capable d'être transformé en différentes configurations selon les tâches et les objets en question. Le ConfigNet présente deux configurations fonctionnelles principales. La première est composée d'un module d'extraction de caractéristiques (FEM), d'un générateur de cartes de décision (DMG) et d'un classificateur ; elle est consacrée à la classification explicative d'images, pour laquelle nous proposons deux structures DMG et un classificateur de mise en commun des moyennes pondérées (WAP) pour les images histopathologiques du cancer du sein. La seconde est une configuration codeur-décodeur consacrée à la segmentation et à la localisation d'objets. Dans cette deuxième configuration, nous proposons un décodeur favorisant l'efficacité et un module de fusion de caractéristiques par éléments (EFFM) guidant la connexion par saut entre l'encodeur et le décodeur pour la détection en ligne des défauts des panneaux d'affichage. En outre, nous développons un module de fusion de caractéristiques guidé par l'attention spatiale et l'attention de canal (SCAFFM) et un décodeur avec une structure de goulot d'étranglement pour la segmentation des tumeurs du sein. Le FEM ou le codeur dans ces deux configurations est construit par apprentissage par transfert à partir de CNNs existants ayant des couches convolutionnelles profondes.

Mots-clés— Apprentissage profond, Réseau neuronal convolutif configurable, Apprentissage faiblement supervisé, Classification explicable, XAI, Segmentation, Cancer du sein, Images histopathologiques, Détection des défauts, Panneau d'affichage

Contents

Acknowledgement **ix**

List of Figures **xi**

List of Tables **xv**

Synthèse en Français de la Thèse **1**

1 General Introduction **33**
1.1 Problem Statement and Objectives 33
1.2 Main Contributions 36
1.3 Organization of Thesis 37

2 Objects and Background **39**
2.1 Breast Cancer Diagnosis 40
2.1.1 Types of Breast Cancer 41
2.1.2 Breast Imaging 45
2.1.3 Datasets for Investigating the CAD System of Breast Cancer Diagnosis 52
2.2 Display Panel Defect Detection 54
2.2.1 Display Panel Manufacturing 54
2.2.2 Online Defect Inspection Algorithms 57
2.2.3 Dataset for Investigating the AOI System of Online Defect Detection of Display Panels 60
2.3 State-of-the-Art Deep Learning Methods 61
2.3.1 Convolutional Neural Networks (CNNs) 61
2.3.2 Weakly-Supervised Learning 71
2.4 Summary 76

3 Configurable Convolutional Neural Networks **77**
3.1 Introduction 78
3.2 Proposed Configurable Network 79
3.3 Breast Cancer Classification with Visual Explanation via the FEM-DMG-Classifier Configuration of ConfigNet 83
3.3.1 Introduction 83
3.3.2 Methodology 84
3.3.3 Experimental Settings 88

3.3.4 Results and Discussion . . . 89
3.3.5 Conclusion . . . 92
3.4 Breast Tumor Segmentation via the Encoder-Decoder Configuration of Our ConfigNet . . . 93
3.4.1 Methodology . . . 93
3.4.2 Experimental Settings . . . 97
3.4.3 Results and Discussion . . . 98
3.4.4 Conclusion . . . 102
3.5 Summary . . . 102

4 ExplaCNet: Explainable Classification of Histopathological Breast Cancer Images Based on Weakly-Supervised Learning 103
4.1 Introduction . . . 104
4.2 Methodology . . . 107
4.2.1 Input Preprocessing . . . 107
4.2.2 Decision Map Generator . . . 107
4.3 Experimental Setting . . . 110
4.3.1 Evaluation Metrics . . . 110
4.3.2 Implementation Details . . . 112
4.4 Results and Discussion . . . 113
4.4.1 Comparison in Performance of Explanation . . . 113
4.4.2 Comparison in Performance of Classification . . . 119
4.4.3 Ablation Study . . . 119
4.5 Conclusions . . . 123

5 EFFNet: Element-Wise Feature Fusion Network for Defect Detection of Display Panels 125
5.1 Introduction . . . 126
5.2 Proposed Method . . . 128
5.2.1 Defect Extraction Module . . . 130
5.2.2 Feature Decoder . . . 131
5.2.3 Element-Wise Feature Fusion Module . . . 132
5.2.4 Loss Function . . . 133
5.3 Experimental Settings . . . 133
5.3.1 Implementation Details . . . 133
5.3.2 Evaluation Metrics . . . 134
5.4 Results and Discussion . . . 134
5.4.1 Ablation Study . . . 134
5.4.2 Comparison with Non-Deep-Learning Methods . . . 135
5.4.3 Comparison with Deep-Learning Methods . . . 138
5.4.4 Analysis of Failure Cases . . . 149
5.5 Conclusion . . . 150

6 General Conclusions and Perspectives 151

List of Publications 155

Bibliography **157**

Acknowledgement

Time passes away day and night, and here I am, reaching the time of graduation. I still remember the excitement, apprehension, and anticipation when I came to INSA for the first time. Although it is unfortunate to be in the time of a global pandemic, I am happy and satisfied during my days at INSA.

First of all, I would like to express sincere gratitude to my supervisor, Prof. Yuemin Zhu. He is the one who opened the door to the field of medical image processing for me. From the decision-making of the research objective, the development of methods, the analysis of experiments, and paper writing and revision to the completion of the Ph.D. project, he has given me guidance and advice with great patience and seriousness. He gave me a very free, open, and pleasant research environment. He provided me with enough soil and nutrients for the progress of my Ph.D. research. He has cultivated my independent research thought and creativity. In addition, he often shared with me anecdotes about French culture and lifestyle, which broadened my horizon. All of these inspired me to establish an optimistic attitude toward life and a rigorous research spirit.

My sincere thanks also go to my co-supervisor, Prof. Zhengjun Liu, for his meticulous guidance and assistance and for enhancing my resilience to stress. Furthermore, I would like to thank him for providing me with the valuable opportunity to study at INSA.

I acknowledge the support from Dr. Yunlong He, Dr. Louise Friot Giroux, and Mehdi Shekarnabi for their great help. I am also grateful for the help of other colleagues, including but not limited to Dr. Yulei Qin, Sophie Carneiro, Samaneh Choopani, and Cyril Malinet. Moreover, I would like to thank Zongze Li, Kelin Wu, Gong Chen, Pujian Guan, Zhonglin Sun, Lijuan Ren, Huiru Ren, and Jiao Zhao, with whom I had fun during my time at INSA.

Finally, I can't thank my parents and family more for their continuous support and encouragement. I am so grateful for their unconditional love and caring.

List of Figures

1 Anatomie du sein (Bazira et al., 2021). 8
2 Différents défauts dans six classes de fonds de panneaux d'affichage. . . 11
3 Architecture de LeNet-5 (LeCun et al., 1998). Chaque plan est une carte de caractéristiques. 12
4 Une illustration de l'architecture d'AlexNet (Krizhevsky et al., 2012), montrant explicitement la délimitation des responsabilités entre les deux GPU. 12
5 Architecture du VGG16 (Ferguson et al., 2017). 13
6 Architecture d'Inception-v1 (a.k.a. GoogLeNet) (Szegedy et al., 2015a). . 13
7 Architecture du FCN (Long et al., 2015). 14
8 Exemples typiques d'images de tumeurs du sein (rangée du haut) et d'images de défauts de panneaux d'affichage (rangée du bas). 17
9 Illustration de ConfigNet 18
10 Illustration de la configuration du classificateur FEM-DMG. 19
11 Illustration de la configuration de l'encodeur-décodeur. 20
12 Une illustration de l'architecture ExplaCNet. 23
13 Défauts typiques difficiles à reconnaître sur les panneaux d'affichage. (a) Le faible défaut dans le fond clair. (b) Le défaut fort dans le fond de couleur sombre. (c) Le défaut de malformation avec une limite ambiguë. Les marques rouges indiquent les régions de défauts. 25
14 Une illustration du modèle EFFNet proposé. 27

2.1 Breast anatomy (Bazira et al., 2021). 40
2.2 Illustration of a mammogram (Wikimedia, 2021). 45
2.3 Normal (left) versus cancerous (right) mammography images (Wikimedia, 2019). 45
2.4 Breast ultrasound examination (source: https://www.radiologyinfo.org/ en/info/breastus). 47
2.5 A. Simple cyst meeting all ultrasound criteria. B. Cluster of cysts (Kossoff, 2000). 47
2.6 Breast magnetic resonance imaging (MRI, source: https://www.mayoclinic.org/diseases-conditions/breast-cancer/diagnosis-treatment/drc-20352475). 48
2.7 Bulky left axillary lymphadenopathy on MRI image. The arrow indicates a 10-mm enhancing retro areolar mass (Shahid et al., 2016). 48

2.8 PET/CT Scan for Breast Scanning (source: https://www.saintjohnscancer.org/breast/breast-health/breast-evaluation/other-tests). 49
2.9 Breast cancer PET/CT images with recurrence in the left and metastases in the right (Zangheri et al., 2004). 50
2.10 Breast biopsy (source: https://www.mayoclinic.org/tests-procedures/breast-biopsy/about/pac-20384812). 51
2.11 histopathological image of invasive (or infiltrating) lobular carcinoma (ILC) obtained during the biopsy (Wikimedia, 2020). 52
2.12 Examples of metastatic (upper row, red regions are metastatic) and normal (bottom row) images from Camelyon16 patch-based dataset. 53
2.13 Slides of breast malignant tumor seen with different magnification factors. 54
2.14 Examples of four benign sub-classes (upper row) and four malignant sub-classes (bottom row) with a magnification factor of 40×. 54
2.15 Array process 55
2.16 Cell process 56
2.17 Module Assembly 57
2.18 Different defects in six classes of backgrounds of display panels. 60
2.19 Architecture of LeNet-5 (LeCun et al., 1998). Each plane is a feature map. 62
2.20 An illustration of the architecture of AlexNet (Krizhevsky et al., 2012), explicitly showing the delineation of responsibilities between the two GPUs. 62
2.21 Architecture of VGG16 (Ferguson et al., 2017). 63
2.22 Architecture of Inception-v1 (a.k.a. GoogLeNet) (Szegedy et al., 2015a). . 63
2.23 Architecture of ResNet-34 (He et al., 2016a). 64
2.24 Architecture of DenseNet (Huang et al., 2017). 64
2.25 Architecture of FCN (Long et al., 2015). 65
2.26 Architecture of U-Net (Ronneberger et al., 2015). 66
2.27 Architecture of SegNet (Badrinarayanan et al., 2017). 67
2.28 Architecture of Attention U-Net (Schlemper et al., 2019). 67
2.29 Illustration of the backpropagation algorithm. 68
2.30 Architecture of CAM (Zhou et al., 2016). 71
2.31 Architecture of WILDCAT (Durand et al., 2017). 72
2.32 Architecture of Grad-CAM (Selvaraju et al., 2017a). 72
2.33 Architecture of Grad-CAM++ (Chattopadhay et al., 2018). 73
2.34 Architecture of ACoL (Zhang et al., 2018a). 74
2.35 Architecture of Attention MIL (Ilse et al., 2018a; Patil et al., 2019). 74
2.36 Architecture of CELNet-CELM (Huang and Chung, 2019). 75
2.37 Architecture of the model proposed by Ciga et al. Ciga and Martel (2021). 75

3.1 Typical examples of breast tumor images (top row) and display panel defect images (bottom row). 79
3.2 Illustration of ConfigNet 80
3.3 Illustration of the FEM-DMG-classifier configuration. 81
3.4 Illustration of the encoder-decoder configuration. 82
3.5 An illustration of MICNet architecture. 85
3.6 The detail of weighted average pooling (WAP) classifier. 87

3.7 Visual explanation of images classification with (a) and without (b) tumors on Camelyon16 patch-based dataset. Red and blue regions in images are tumors and healthy tissues, respectively. Images inside the black, red and green boxes are original images, the ground truth, and explanation map results from our model, respectively. 91
3.8 Illustration of SCAFFNet. 94
3.9 The detail of SCAFFM architecture. 96
3.10 P-R curve of the comparison. 100
3.11 ROC curve of the comparison. 100
3.12 Some visual results of the comparison. 101

4.1 An illustration of ExplaCNet architecture. 108
4.2 The architecture of decision map generator (DMG) module. “Same" means that the convolution keeps the resolution. “2" denotes that the output has two channels. 109
4.3 Top row: Confusion matrix over all pixels of Camelyon16 patch-based test set. Bottom row: MPA versus MPE. 115
4.4 Examples of explanation maps for tumor input. Red and blue regions represent the presence and absence of metastatic tissues, respectively, which provide evidence for the classification decision. 116
4.5 Examples of explanation maps for normal input. Red and blue regions represent the presence and absence of metastatic tissues, respectively, which provide evidence for the classification decision. 117
4.6 Top row: Confusion matrix over all pixels of Camelyon16 patch-based test set. Bottom row: MPA versus MPE. 121
4.7 Examples of explanation maps for tumor input in ablation study. Red and blue regions represent the presence and absence of metastatic tissues, respectively, which provide evidence for the classification decision. 121
4.8 Examples of explanation maps for normal input in ablation study. Red and blue regions represent the presence and absence of metastatic tissues, respectively, which provide evidence for the classification decision. 122

5.1 Typical hard-to-recognize defects in display panels. (a) The weak defect in the light-colored background. (b) The strong defect in the dark-colored background. (c) The malformation defect with an ambiguous boundary. The red marks indicate the regions of defects. 127
5.2 An illustration of the proposed EFFNet model. 129
5.3 The architecture of element-wise feature fusion module. 132
5.4 Histogram of the evaluation results of different methods. 136
5.5 Comparison of the proposed method with traditional image processing methods. Ori denotes the original image, and GT means the ground truth. The numbers at the bottom of the images represent the corresponding mIoU values of the detection, where the values in red are the best, and the ones in blue are the second-best. 137
5.6 Learning curves of CNN-based methods. (a) Training loss. (b) Validation loss. The curves on the right are the enlargements of the corresponding areas of the left learning curves. 138

5.7 Evaluation results of our method and other CNNs. (a) TDR-mIoU curve. (b) P-R curve. (c) Histogram of the evaluation metrics. 140
5.8 The visual comparison of our method with other state-of-the-art segmentation CNNs. Ori means original image, and GT represents the ground truth. AU-Net is Attention U-Net. The numbers underneath the images are the corresponding mIoU values of the detection, where the values in red are the best, and the ones in blue are the second best. . . . 141
5.9 Evaluation curves of CNN-based methods trained with two sizes of datasets. (a) and (c) TDR-mIoU curve. (b) and (d) P-R curve. 142
5.10 Evaluation curves of CNN-based methods in the robustness experiment. (a) and (c) TDR-mIoU curve. (b) and (d) P-R curve. 144
5.11 TDR-mIoU curves of CNN-based methods in the study of different backgrounds. 146
5.12 P-R curves of CNN-based methods in the study of different backgrounds. 147
5.13 Failures of the proposed method. 149

List of Tables

2.1 Distribution of training and test sets after data augmentation. 61

3.1 Classification results (in %) on BreakHis test set. The best performance is shown in bold. N/A denotes the value that was not provided in the original work. 90
3.2 Encoder structure. 93
3.3 Decoder structure. 95
3.4 Comparison results (in %) of segmentation performance of different methods on Camelyon16 patch-based test set. The best performance is indicated in bold. Sen. and Spe. are simplified from Sensitivity and Specificity, respectively. 99

4.1 Comparison results (in %) of explanation and classification performance on Camelyon16 patch-based test set. The best performance is indicated in bold. 114
4.2 Classification results (in %) on BreakHis test set. The best performance is indicated in bold. 118
4.3 Ablation study (in %) on Camelyon16 patch-based test set. The best performance is indicated in bold. 120

5.1 Defect Extraction Module. 130
5.2 Evaluation results of ablation study in the effects of EFFM. 134
5.3 Evaluation results of our encoder with five transfer learning strategies. . 135
5.4 Evaluation results of our method and other non-deep-learning methods. 136
5.5 Quantitative comparison of deep-learning methods. 139
5.6 Evaluation results of deep-learning methods in the study of dataset size. 143
5.7 Evaluation results of deep-learning methods in the study of motion blur noise. 145
5.8 Evaluation results of deep-learning methods in the study of different backgrounds. 148

Synthèse en Français de la Thèse

Chapitre 1 Introduction Générale

Énoncé du Problème et Objectifs

Les principaux objectifs de cette thèse sont de deux ordres: L'un est de développer des modèles d'apprentissage profond intrinsèquement explicables pour les systèmes de diagnostic assisté par ordinateur (DAO) afin de résoudre le problème de la classification histopathologique des images du cancer du sein et de fournir un support fiable pour le déploiement de systèmes DAO basés sur l'apprentissage profond dans des contextes cliniques réels. Un autre objectif est de développer des modèles d'apprentissage profond très efficaces et précis pour les systèmes d'inspection optique automatique (IOA) afin de détecter en ligne les défauts des panneaux d'affichage sur les chaînes de montage des usines.

Selon son site, le cancer du sein comprend les cancers invasifs et non invasifs. Le cancer du sein invasif (Harris et al., 2016) survient lorsque des cellules malades situées à l'intérieur des lobules ou des canaux lactifères se détachent à proximité du tissu mammaire. Il se compose de nombreux sous-types différents, comme le carcinome lobulaire infiltrant (CLI) (Yedjou et al., 2022), le carcinome canalaire infiltrant (CCI) (Zhou et al., 2021), le carcinome médullaire (Mateo et al., 2017), le carcinome mucineux (Komenaka et al., 2004), Carcinome tubulaire (Huang et al., 2021a), Cancer du sein inflammatoire (Hu et al., 2021), Maladie de Paget du sein (Sakorafas et al., 2001), Tumeur phyllode (PARK et al., 2021), et Cancer du sein triple négatif (Li et al., 2019b). Le cancer du sein non invasif (West et al., 2017), quant à lui, ne s'étend pas à partir des lobules ou des canaux où il est situé. Il présente principalement deux sous-types, à savoir le carcinome canalaire in situ (CCIS) (Weedon-Fekjær et al., 2021) et le carcinome lobulaire in situ (CLIS) (Masannat et al., 2013) où "in situ" représente "en place". La complexité de son sous-type, son occurrence la plus répandue et la négligence des femmes en matière d'auto-examen des seins et d'examen clinique en font le cancer le plus meurtrier pour les femmes atteintes de cancer dans le monde, avec un ratio de 1/4 parmi tous les cas de cancer et 1/6 de décès par cancer (685,000 décès en 2020) (Akram et al., 2017; Sung et al., 2021).

Selon l'Organisation mondiale de la santé (OMS), le fondement de la réglementation du cancer du sein visant à améliorer les résultats du traitement et la survie réside dans son diagnostic précoce. Grâce au développement de la technologie de l'imagerie médicale, la réglementation peut récemment être réalisée selon de nombreuses approches répondant à des exigences diverses. La mammographie (Gøtzsche and Jørgensen, 2013)

est une technique standard de dépistage de masse qui permet d'obtenir des images des os, des tissus mous et des vaisseaux sanguins en même temps. L'échographie (Ozmen et al., 2015) est une technique rentable et sans radiation qui permet d'identifier les kystes et les masses solides, et elle a été recommandée comme complément pour évaluer les masses découvertes par la mammographie. L'imagerie par résonance magnétique (IRM) (Kuhl, 2019) produit des images détaillées à différentes sections transversales grâce à de puissants champs magnétiques et à des ondes radio générées par ordinateur, ce qui la rend capable de détecter de petits détails des tissus mous qui ne peuvent être détectés par la mammographie. La tomographie par ordinateur (TO) (Boone et al., 2006) scanne une série d'images à rayons X sous différents angles, puis utilise des algorithmes de reconstruction tomographique pour produire des images en coupe transversale (tranches) des os, des vaisseaux sanguins et des tissus mous. La tomographie par émission de positons (TEP) (Vercher-Conejero et al., 2015) visualise et mesure les changements dans les processus métaboliques et autres activités physiologiques (tels que le flux sanguin, la composition chimique régionale et l'absorption) en injectant des substances radioactives (radionucléides) dans une veine périphérique. Bien que les techniques d'imagerie médicale susmentionnées soient non invasives et présentent des avantages spécifiques, elles souffrent de risques de radiation, de coûts élevés ou de faibles sensibilités : (Onega et al., 2016; Hooley et al., 2013; Roganovic et al., 2015). La biopsie (Soo et al., 2019) est le seul moyen définitif et "l'étalon-or" pour diagnostiquer le cancer du sein. Elle consiste à prélever un morceau de tissu ou un échantillon de cellules dans le corps afin de l'examiner au microscope. Cependant, la procédure de diagnostic par biopsie prend beaucoup de temps et nécessite des experts humains bien expérimentés qui ont besoin d'années de formation. Heureusement, les systèmes de diagnostic assisté par ordinateur (DAO) basés sur l'apprentissage profond (Qiu et al., 2017; Cong et al., 2020; Meng et al., 2021) qui ont émergé ces dernières années ont le potentiel d'alléger efficacement la plupart des charges de travail des pathologistes, améliorant ainsi considérablement ce problème.

Actuellement, les méthodes d'apprentissage profond ont fait de grands progrès et obtenu des résultats impressionnants dans la reconnaissance d'objets et le traitement d'images (Chen et al., 2017; Voulodimos et al., 2018; Tian et al., 2020; Esteva et al., 2021; Zheng et al., 2021; Dong et al., 2022). Les réseaux de neurones convolutifs (RNCs), en tant que produits les plus influents, sont largement utilisés pour la classification d'images et la localisation de cibles en raison de leur capacité unique à apprendre des caractéristiques complexes, comme VGG (Simonyan and Zisserman, 2014a), ResNet (He et al., 2016a), WILDCAT (Durand et al., 2017), et ACoL Zhang et al. (2018a). Cela a inspiré de nombreux chercheurs (Spanhol et al., 2016a; Bayramoglu et al., 2016a; Sun et al., 2021; Song et al., 2017; Sudharshan et al., 2019; Kumar et al., 2020; Gour et al., 2020; Schirris et al., 2022; Xu et al., 2019a) à appliquer cette méthode pour étudier la classification des images histopathologiques du cancer du sein (c'est-à-dire la procédure clé et le but ultime de la biopsie mammaire). Bien que les méthodes d'apprentissage profond aient fait d'énormes progrès dans ce domaine et aient même obtenu des résultats équivalents à ceux des experts humains, la plupart des études n'ont pas tenu compte des explications logiques ou rationnelles (c'est-à-dire qu'elles n'ont pas fourni d'explicabilité) pour la décision du réseau. L'introduction de l'explicabilité dans les modèles d'apprentissage profond est un moyen essentiel de débloquer leur

nature de "boîte noire" qui entrave leur transparence, indispensable pour la sécurité, l'éthique, la compréhension humaine et la fiabilité du diagnostic clinique. Le manque d'explicabilité a empêché le déploiement de systèmes de DAO basés sur l'apprentissage profond dans des contextes cliniques réels, ce qui conduit au premier objectif de cette thèse et indique sa grande importance.

De même, en tant que composant d'affichage de presque tous les appareils électroniques modernes (tels que les smartwatches, les téléphones cellulaires, les ordinateurs portables, les commandes centrales des voitures et les téléviseurs LCD), l'écran à cristaux liquides à transistors à couches minces (TFT-LCD) et la diode électroluminescente organique (OLED) font l'objet d'une demande considérable dans le monde entier. Cependant, la fabrication des écrans LCD et OLED est fastidieuse et compliquée. Prenons l'exemple de l'écran LCD. Son processus de production est principalement divisé en trois parties : le processus de matrice de la section avant, l'assemblage du panneau de la section intermédiaire (cellule) et l'assemblage du module de la section finale (module). Parmi ces opérations, le processus de la section avant comprend le nettoyage, la formation du film, le revêtement de la résine photosensible, l'exposition, le développement, la gravure et le décapage. Bien que le processus de fabrication soit généralement réalisé dans une salle blanche, divers défauts tels que la poussière, la saleté, les courts-circuits et les ruptures se produisent inévitablement à la surface du produit pour des raisons techniques, ce qui entraîne des produits défectueux. L'apparition de ces produits défectueux affecte sérieusement le rendement des écrans LCD et OLED, et donc directement la rentabilité. Par conséquent, la surveillance et l'inspection efficaces de ce processus de fabrication pendant la fabrication et la réparation ou l'élimination en temps voulu des produits défectueux avant l'emballage sont essentielles pour réduire le gaspillage des ressources, réduire le coût du temps et améliorer le taux de rendement.

Trois types de méthodes de détection des défauts des panneaux ont été adoptés actuellement : l'inspection visuelle humaine, l'inspection électrique et l'inspection optique automatisée (IOA). L'inspection visuelle humaine est la méthode traditionnelle et primitive de détection des défauts, qui est limitée sur deux aspects, à savoir une faible précision de détection des défauts observés à l'œil nu et une détection inefficace des défauts observés par échantillonnage microscopique. En outre, elle est subjective, incertaine et facile à mal juger, et le stockage et l'interrogation des données d'inspection sont peu pratiques. Cette méthode d'inspection ne peut pas répondre aux besoins de la production de masse et rapide. Les méthodes de détection électrique, telles que le balayage par sonde, le couplage photoélectrique et la détection des circuits conducteurs, ne peuvent être utilisées que pour les défauts fonctionnels causés par des facteurs électriques et ne conviennent pas à la détection dans le processus de fabrication. L'IOA (He and Sun, 2015; Yuan et al., 2015; Wang et al., 2018; Tsai and Hung, 2005; Tsai et al., 2007) basée sur la vision par ordinateur est la méthode de détection des défauts de surface qui connaît la croissance la plus rapide et qui est la plus utilisée en raison de sa nature sans contact, de son automatisation élevée et de son évolutivité. Cependant, les techniques traditionnelles de vision par ordinateur, telles que le détecteur d'Otsu (He and Sun, 2015), le détecteur de bords de Canny (Wang et al., 2018), la transformée de Fourier (Tsai and Hung, 2005; Tsai et al., 2007), les filtres gaussiens et de Gabor (Tong et al., 2016), le flux optique (Tsai et al., 2011), et

la machine à vecteurs de support (SVM) (Chu et al., 2017), reconnaissent les défauts principalement en se basant sur des caractéristiques et des règles fabriquées à la main qui sont limitées dans la représentation des caractéristiques. Elles peuvent difficilement répondre aux exigences de la détection industrielle en ligne de panneaux d'affichage présentant des défauts à faible contraste, un bruit de fond complexe et diversifié, et des limites ambiguës. En revanche, les méthodes d'apprentissage profond actuellement développées (Liu et al., 2020a; Hu and Wang, 2020; Zou et al., 2018; Song et al., 2020; Lee et al., 2022; Li and Li, 2021; Chang et al., 2022; Yao and Li, 2022; Li and Wang, 2022) avec des capacités supérieures d'extraction de caractéristiques et d'apprentissage et des succès considérables dans la reconnaissance d'objets apportent une solution à ce problème, ce qui a inspiré le deuxième objectif de cette thèse.

En général, l'architecture de réseau d'un modèle d'apprentissage profond est fixée pour une tâche cible spécifique (par exemple, la segmentation, la localisation ou la classification) et un objet d'investigation (par exemple, la classification histopathologique d'une image de cancer du sein ou la détection de défauts sur un écran). Par exemple, pour les tâches de segmentation ou de localisation, où le but est d'obtenir des résultats au niveau des pixels (c'est-à-dire des masques binaires), il est nécessaire que le modèle de réseau soit capable d'extraire le contour approximatif et les informations de localisation de la cible. Les principales architectures de réseau actuelles de cet aspect consistent principalement en des structures d'encodeur-décodeur qui permettent une sortie de bout en bout (c'est-à-dire que la résolution d'entrée est égale à la résolution de sortie), comme le réseau U-Net (Ronneberger et al., 2015) et sa série (Alom et al., 2018; Schlemper et al., 2019; Thomas et al., 2020; Zunair and Hamza, 2021; Lin et al., 2022), SegNet (Badrinarayanan et al., 2017), DeepCrack (Zou et al., 2018), et EDRNet (Song et al., 2020). Pour les tâches de classification, le modèle de réseau doit reconnaître la catégorie de la cible dans l'image (par exemple, bénigne ou maligne), ce qui nécessite qu'il soit capable d'extraire les informations les plus discriminantes pertinentes pour la région cible. Les réseaux de classification actuels sont principalement composés d'un module d'extraction de caractéristiques (c'est-à-dire des couches convolutionnelles complètes avec une certaine profondeur et largeur) et d'un module de classification (par ex, couches entièrement connectées ou couches linéaires), tels que VGG (Simonyan and Zisserman, 2014a), ResNet (He et al., 2016a), Inception (Szegedy et al., 2015a, 2016, 2017), DenseNet (Huang et al., 2017) et ShuffleNet (Zhang et al., 2018b; Ma et al., 2018). Ainsi, la plupart des modèles d'apprentissage profond existants sont orientés vers la tâche (c'est-à-dire qu'ils manquent de généralité), ce qui est préjudiciable à l'étude et à la mise en œuvre de notre travail. Sur la base d'une étude complète des travaux précédemment rapportés, nous avons proposé un cadre d'apprentissage profond appelé réseau configurable (ConfigNet) qui peut être configuré de manière adaptative pour différentes tâches et objets. Il peut être appliqué pour réaliser à la fois une classification explicable des images histopathologiques du cancer du sein et une détection en ligne des défauts des panneaux d'affichage.

Principales Contributions

Les principales contributions de cette thèse sont détaillées comme suit:

- Réseaux neuronaux convolutifs configurables (Chapitre 3).

Nous avons proposé un réseau neuronal convolutif configurable (ConfigNet) capable de se transformer en différentes configurations qui s'adaptent à plusieurs tâches et objets. Le ConfigNet possède principalement deux configurations fonctionnelles définies comme la configuration FEM-DMG-classifieur pour la classification d'images explicables et la configuration encodeur-décodeur pour la segmentation et la localisation d'objets. La configuration FEM-DMG-classifieur est composée d'un module d'extraction de caractéristiques (FEM), d'un générateur de cartes de décision (DMG) et d'un classifieur. Les squelettes du FEM et du codeur de ces deux configurations sont tous deux construits par l'apprentissage par transfert de CNN existants avec des couches convolutives profondes. Pour la configuration FEM-DMG-classifieur (MICNet) basée sur VGG11, un DMG de base et un classifieur WAP (weighted average pooling), des expériences approfondies sur les ensembles de données basés sur les patchs BreakHis et Camelyon16 démontrent qu'elle surpasse les autres modèles CNN dans la classification des images histopathologiques du cancer du sein et peut fournir une explication visuelle logique qui soutient la prédiction du réseau. En ce qui concerne la configuration codeur-décodeur (SCAFFNet) basée sur un module de fusion de caractéristiques guidé par l'attention spatiale et de canal (SCAFFM) et un décodeur avec une structure de "goulot d'étranglement", les résultats expérimentaux sur le jeu de données Camelyon16 basé sur des patchs montrent qu'il surpasse les modèles de segmentation de pointe dans la segmentation des tumeurs du sein, en particulier dans le cas difficile où les tumeurs du sein ont des limites complexes.

- ExplaCNet: Classification explicable d'images histopathologiques du cancer du sein basée sur un apprentissage faiblement supervisé (Chapitre 4).

  Nous avons proposé un nouveau réseau basé sur l'apprentissage faiblement supervisé, ExplaCNet, pour réaliser la classification explicable des images histopathologiques du cancer du sein. Nous avons utilisé l'apprentissage par instances multiples (MIL) pour encourager l'ExplaCNet à identifier davantage de tissus normaux et un classificateur adaptatif de mise en commun de la moyenne pondérée (WAP) qui fusionne plusieurs instances en une probabilité de sac d'instances pour forcer l'ExplaCNet à apprendre à reconnaître davantage de tissus de lésions. En particulier, nous avons développé un générateur de cartes de décision (DMG) avec des filtres multi-échelles qui permettent un compromis raisonnable dans la capacité de l'ExplaCNet à identifier les tissus normaux et lésionnels afin de générer des cartes de décision raffinées pour la classification finale et l'explication visuelle. Les résultats expérimentaux sur les jeux de données Camelyon16 et BreakHis ont démontré que notre ExplaCNet surpasse les méthodes explicatives de l'état de l'art en termes d'explication visuelle tout en conservant une performance de classification compétitive.

- EFFNet: Réseau de fusion de caractéristiques par éléments pour la détection des défauts des panneaux d'affichage (Chapitre 5).

  Nous avons mis au point un nouveau réseau de fusion de caractéristiques par éléments (EFFNet) afin de détecter avec une grande précision et en temps réel les défauts des panneaux d'affichage. La méthode a adopté un apprentissage par

transfert et par réglage fin pour les couches d'extraction de caractéristiques et un décodeur avec une complexité de calcul relativement faible. En particulier, un module de fusion des caractéristiques basé sur l'addition élément par élément des caractéristiques pyramidales a été proposé dans la connexion par saut saut pour améliorer l'efficacité et la précision de la détection. Notre méthode a été comparée aux techniques traditionnelles de détection des défauts et aux modèles RNC les plus avancés. En outre, les effets de la taille de l'ensemble de données d'entraînement, du flou de mouvement et de différents fonds sur les performances de la méthode proposée ont été étudiés. Des expériences approfondies démontrent que le réseau développé peut détecter avec précision des défauts à textures complexes, à limites ambiguës et à faible contraste. Il présente également une bonne robustesse face au flou de mouvement. Il surpasse les méthodes les plus récentes en termes de mIoU, MPA et mesure F1. De plus, elle est capable de détecter des défauts à des vitesses allant jusqu'à 159 fps/s avec des images d'entrée de taille 256×256.

Organisation de la Thèse

Le manuscrit de la thèse est organisé comme suit:

Au Chapitre 2, intitulé "Objets et Contexte", nous présentons en détail le diagnostic du cancer du sein, notamment les types de cancer du sein, les techniques d'imagerie mammaire et deux ensembles de données publiques d'images histopathologiques pour l'étude des algorithmes de systèmes de diagnostic assistés par ordinateur. Nous présentons également la détection des défauts des panneaux d'affichage, notamment la fabrication des panneaux d'affichage, l'inspection en ligne des défauts et le jeu de données d'images de défauts de panneaux d'affichage pour l'étude des algorithmes de systèmes d'inspection optique automatisés. En outre, nous décrivons deux méthodes d'apprentissage profond étroitement liées à cette thèse. L'une est les réseaux de neurones convolutifs, et l'autre l'apprentissage faiblement supervisé.

Au Chapitre 3, intitulé "Réseaux Neuronaux Convolutifs Configurables", nous avons proposé un réseau neuronal convolutif configurable appelé ConfigNet, qui peut être configuré de manière adaptative pour différentes tâches et différents objets afin de réaliser la classification explicable d'images histopathologiques du cancer du sein et la détection en ligne des défauts des écrans. Le chapitre présente d'abord les grandes lignes du ConfigNet développé. Ensuite, les performances de la configuration de son classificateur FEM-DMG (appelé MICNet) pour réaliser la classification d'images histopathologiques du cancer du sein tout en fournissant des explications visuelles logiques sont étudiées. L'architecture complète de notre MICNet est décrite en détail, y compris la stratégie MIL, le FEM et le classificateur WAP (weighted average pooling), ainsi que la manière dont il génère des explications. La performance de notre MICNet a été évaluée sur deux jeux de données disponibles publiquement. Enfin, la configuration de l'encodeur-décodeur (appelée SCAFFNet) basée sur le module de fusion de caractéristiques guidé par l'attention spatiale et l'attention de canal (SCAFFM) est comparée à d'autres modèles de segmentation de pointe sur le jeu de données basé sur les patchs de Camelyon16 afin d'évaluer les performances de segmentation, où le SCAFFNet est présenté en détail.

Au Chapitre 4, intitulé " ExplaCNet : ExplaCNet : Explainable Classification of Histopathological Breast Cancer Images Based on Weakly-Supervised Learning ", nous avons développé une configuration de classificateur FEM-DMG favorisant l'explicabilité (appelée ExplaCNet) de notre ConfigNet afin de résoudre le conflit entre les cartes d'explication mieux compréhensibles par l'homme et les décisions de classification plus précises et de parvenir à une classification cliniquement explicable des images histopathologiques du cancer du sein. La nouvelle conception du DMG avec des filtres multi-échelles, permettant un compromis raisonnable dans le conflit et améliorant l'explicabilité tout en conservant la performance de classification, a été détaillée. En outre, une étude d'ablation a été réalisée pour démontrer l'efficacité des composants clés de notre ExplaCNet. Ce dernier a été évalué en termes d'explication et de classification sur deux ensembles de données sur le cancer du sein, avec de nombreuses méthodes d'apprentissage profond de pointe.

Au Chapitre 5, intitulé "EFFNet: Réseau de Fusion de Caractéristiques par éléments Pour la Détection des Défauts des Panneaux d'affichage", nous avons proposé une configuration de codeur-décodeur favorisant l'efficacité (appelée EFFNet, c'est-à-dire réseau de fusion de caractéristiques par éléments) de notre ConfigNet pour la détection en ligne des défauts des panneaux d'affichage. L'EFFNet contient un module d'extraction des défauts, un décodeur de caractéristiques et un module de fusion des caractéristiques par éléments (EFFM), qui ont été présentés en détail. Des études d'ablation sur l'EFFM et différentes stratégies d'apprentissage par transfert ont été mises en œuvre. Des expériences approfondies, y compris des comparaisons avec des méthodes d'apprentissage non approfondi et d'apprentissage approfondi et une discussion sur les effets de la taille de l'ensemble de données d'entraînement, du flou de mouvement et des arrière-plans, ont été menées pour démontrer la supériorité et la robustesse de notre EFFNet.

Au chapitre 6, intitulé "Conclusions Générales et Perspectives", nous résumons le contenu principal de cette thèse, y compris nos contributions majeures, nos conclusions générales et nos perspectives futures.

Chapitre 2 Objets et Contexte

Diagnostic du Cancer du Sein

Le sein (comme le montre la Fig. 1) est une glande tubulo-alvéolaire entourée de tissu adipeux gras (y compris du tissu conjonctif fibreux et des ligaments) contenant un réseau de nerfs, de vaisseaux sanguins, de vaisseaux lymphatiques et de ganglions lymphatiques (Stingl et al., 2005; Delotte et al., 2009; Fahad Ullah, 2019). En particulier, les seins féminins se composent normalement de 12 à 20 lobes qui sont subdivisés en de nombreux lobules plus petits, et ils sont reliés par des canaux lactifères (Tanis et al., 2001). La structure de la glande mammaire contient généralement un épithélium stratifié, composé de myoépithélium et d'épithélium, délimité par une membrane basale et ancré dans un gabarit de cellules vasculaires, lymphatiques et stromales (Stingl et al., 2006). Le cancer du sein survient pour les raisons suivantes : 1) Le système immunitaire humain est incapable d'identifier les dommages causés à l'acide désoxyribonucléique

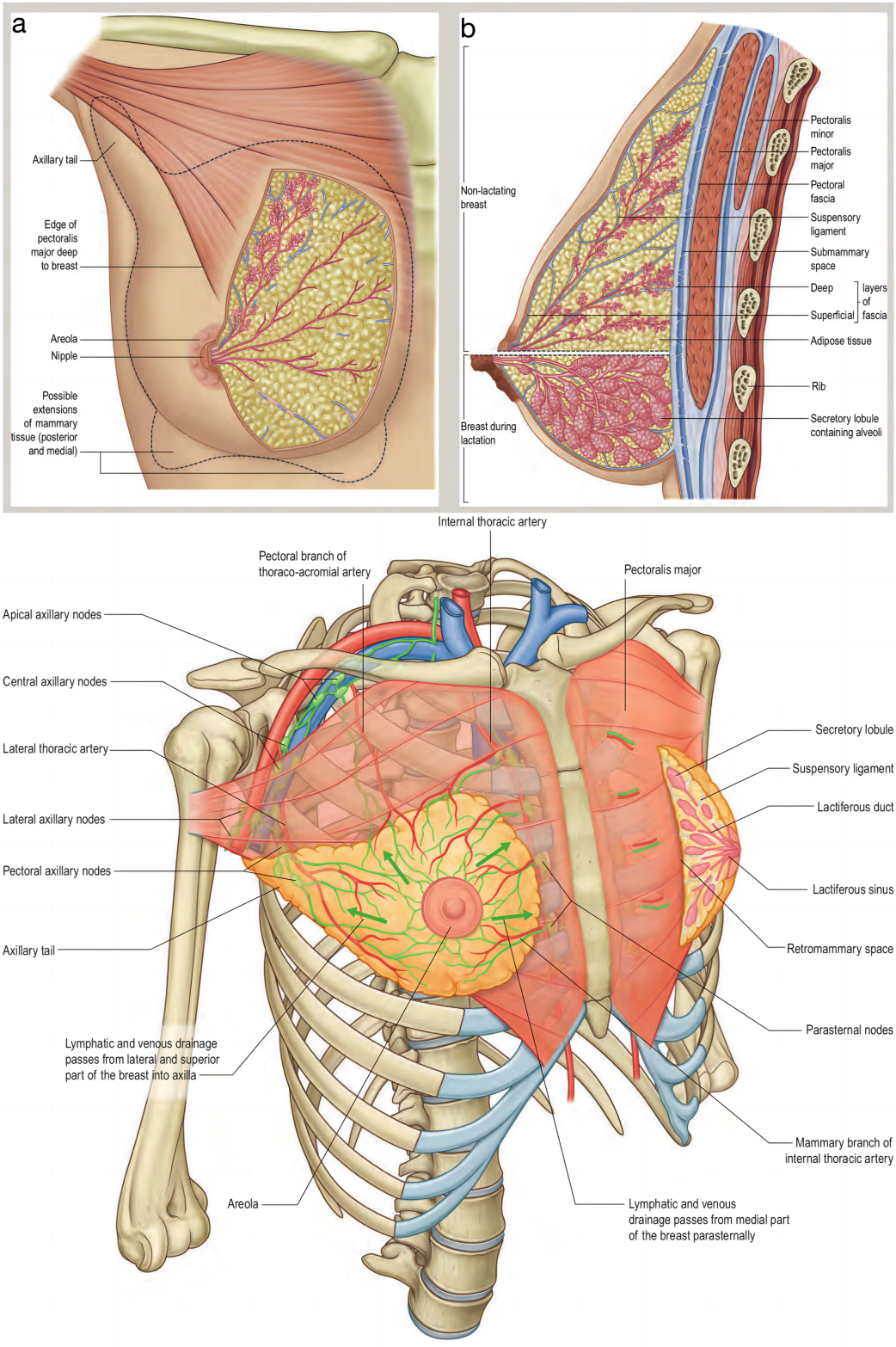

Figure 1: Anatomie du sein (Bazira et al., 2021).

(ADN), les erreurs dans les gènes tels que P53, BRCA1 et BRCA2, et les altérations héréditaires et de les résoudre, ce qui entraîne des modifications tumorigènes dans les

cellules myoépithéliales ou épithéliales (ou les cellules souches ayant la capacité de se développer en cellules myoépithéliales ou épithéliales), qui à leur tour conduisent au cancer ; 2) les voies RAS/MEK/ERK et PI3K/AKT protègent les cellules normales du suicide cellulaire. Lorsque les gènes impliqués dans le codage de ces voies de protection sont mutés, les cellules ne parviennent pas à se suicider lorsqu'elles n'en ont plus besoin, ce qui entraîne le développement de cancers associés à l'exposition aux œstrogènes. Dans les cellules cancéreuses, la télomérase inverse le raccourcissement des chromosomes et permet une réplication cellulaire extensive (Hanahan and Weinberg, 2000). Alors que les cellules tumorales obtiennent généralement leur approvisionnement en nutriments et en oxygène par le biais de l'angiogenèse, les cellules cancéreuses franchissent cette limite et peuvent pénétrer dans la circulation sanguine, le tissu lymphatique et d'autres tissus de l'organisme pour produire des tumeurs secondaires (Jain, 2005; Gupta and Massagué, 2006).

La mammographie est une technique d'imagerie qui utilise une faible dose de rayons X (environ 30 kVp) pour examiner le sein humain. Normalement, les grosseurs cancéreuses et les dépôts de calcium apparaissent plus brillants sur les mammographies, ce qui permet d'établir un diagnostic et de procéder à un dépistage. L'échographie (aussi appelée ultrasonographie) est une technique d'imagerie non invasive qui utilise des ondes sonores à haute fréquence (7,5-15 MHz (Kossoff, 2000)) sans radiation et les échos correspondants pour visualiser l'intérieur des organes concernés. Elle permet de visualiser certains changements dans le sein, tels que les kystes remplis de liquide et le flux sanguin dans la région du sein, qui sont peu susceptibles d'être observés sur une mammographie. L'IRM du sein est une technique d'imagerie qui utilise des aimants et des ondes radio pour produire des images tridimensionnelles et détaillées de l'anatomie du sein en se basant sur la propriété magnétique du noyau d'hydrogène qui est abondant dans l'eau qui constitue les tissus vivants. La tomographie par émission de positons (plus connue sous le nom de PET-CT ou PET/CT) est une technique d'imagerie de médecine nucléaire qui combine un scanner de tomographie par émission de positons (PET) et un scanner de tomographie par rayons X (CT) dans un portique. Elle génère une séquence d'images provenant des deux appareils au cours de la même séance et les combine en une seule image superposée (co-registrée). Cette combinaison permet d'aligner (ou de corréler) la distribution spatiale de l'activité métabolique ou biochimique du corps obtenue par TEP avec les caractéristiques anatomiques obtenues par TDM. La biopsie mammaire est une procédure d'examen du cancer du sein au cours de laquelle un petit échantillon de tissu mammaire est prélevé par des méthodes chirurgicales telles que l'excision locale, l'aspiration par ponction à l'aiguille, le grattage et l'enlèvement, puis sectionné en vue d'un examen microscopique.

Cette thèse s'est principalement concentrée sur l'investigation de l'algorithme du système DAO pour le diagnostic du cancer du sein avec biopsie, qui va finalement à la classification des images histopathologiques du cancer du sein. Afin d'évaluer et de vérifier l'efficacité de notre algorithme (c'est-à-dire les modèles d'apprentissage profond) dans une variété de tâches d'histopathologie numérique du cancer du sein cliniquement pertinentes, deux ensembles de données publiques ont été utilisés dans nos expériences. Le premier est le jeu de données basé sur le patch Camelyon16 (Rony et al., 2019) qui est dérivé du jeu de données Camelyon16 (Bejnordi et al., 2017), et le

second est le jeu de données BreakHis (Spanhol et al., 2016b).

Détection des Défauts du Panneau d'affichage

La fabrication d'un panneau d'affichage comprend le processus de la matrice, le processus de la cellule et le processus d'assemblage du module. Le procédé Array du panneau LCD comprend principalement le film, la lumière jaune, la gravure et le pelage. Comme les électrons sont nécessaires pour piloter le mouvement et l'alignement des molécules de l'écran LCD, des pièces conductrices qui les contrôlent sont nécessaires sur le verre TFT (le support de l'écran LCD). L'oxyde d'étain et d'indium (ITO) est l'un des matériaux qui peuvent être utilisés à cette fin. L'ITO est transparent et peut donc agir comme un cristal conducteur en couche mince qui ne bloque pas le rétroéclairage. Comme pour l'impression du circuit sur la carte de circuit imprimé, le film ITO nécessite de dessiner le circuit conducteur sur l'ensemble de la carte LCD. Le panneau LCD est structuré comme un sandwich, avec le verre TFT en dessous et le filtre coloré au-dessus. Ainsi, le processus de la cellule du terminal comprend le collage du verre TFT sur le dessus et du filtre coloré sur le dessous. Le processus d'assemblage du module consiste principalement en l'ajustement par pression du circuit intégré de commande et du substrat LCD et en l'intégration de la carte de circuit imprimé. Ce processus permet de transmettre le signal d'affichage reçu du circuit de commande principal au circuit intégré de commande afin de faire tourner les molécules LCD et d'afficher l'image. En outre, la partie rétroéclairage sera intégrée au substrat LCD à ce stade, et le panneau LCD est finalement terminé.

Les panneaux LCD et OLED présentent des surfaces périodiquement texturées au stade de la matrice. Les méthodes traditionnelles telles que la détection des bords et la segmentation par seuil ne sont pas efficaces pour détecter les défauts dans de tels arrière-plans où la valeur du pixel du défaut est proche de celle de la texture de fond. Les méthodes actuelles d'inspection optique (c'est-à-dire les algorithmes IOA) pour la fabrication des panneaux d'affichage sont principalement de trois types : différentiel, transformation et statistique. La méthode différentielle est une méthode de détection simple et directe. Son idée principale est d'obtenir l'image résiduelle en effectuant l'opération différentielle entre l'image modèle sans défaut et l'image à détecter, puis de déterminer les défauts selon des algorithmes de segmentation. La méthode comprend principalement la génération d'images modèles sans défaut, l'alignement des modèles et des images de défauts, l'extraction et la détermination des défauts. De nombreux travaux ont été réalisés pour améliorer les performances de cette méthode. La méthode de transformation transforme d'abord le signal d'image du domaine spatial au domaine spectral en utilisant la transformée de Fourier, la transformée en ondelettes, la transformée de Gabor ou la transformée en cosinus discrète. Ensuite, les composantes de fréquence de l'arrière-plan de texture répétitive sont filtrées dans le domaine spectral, et l'information anormale locale des défauts est conservée. Ensuite, la reconstruction de l'image est effectuée par la transformée inverse correspondante pour obtenir une image sans texture répétitive. Enfin, la détermination des défauts de l'image est effectuée par segmentation de l'image. La méthode statistique effectue une extraction de caractéristiques ou une réduction de la dimensionnalité des données sur la base des données d'image acquises. Les informations extraites, telles que les carac-

téristiques de texture, les caractéristiques géométriques et les caractéristiques d'échelle de gris, sont ensuite introduites dans le classificateur pour compléter la détermination de la présence ou de l'absence de défauts ou la reconnaissance des classes de défauts. Il existe de nombreux travaux de recherche sur l'extraction de caractéristiques et la construction de classificateurs.

Nous avons utilisé un ensemble de données sur les défauts d'un panneau d'affichage collectées par un système de microscope dans le cadre d'une fabrication industrielle réelle pour évaluer et vérifier l'efficacité de notre méthode. L'ensemble de données contient 571 images défectueuses de 1024×768 pixels, et les images présentent six classes de fonds. Chacune d'entre elles présente différents types de défauts, tels que le corps étranger, le film décollé, l'adhésif du film, la malformation, etc. comme le montre la Fig. 2.

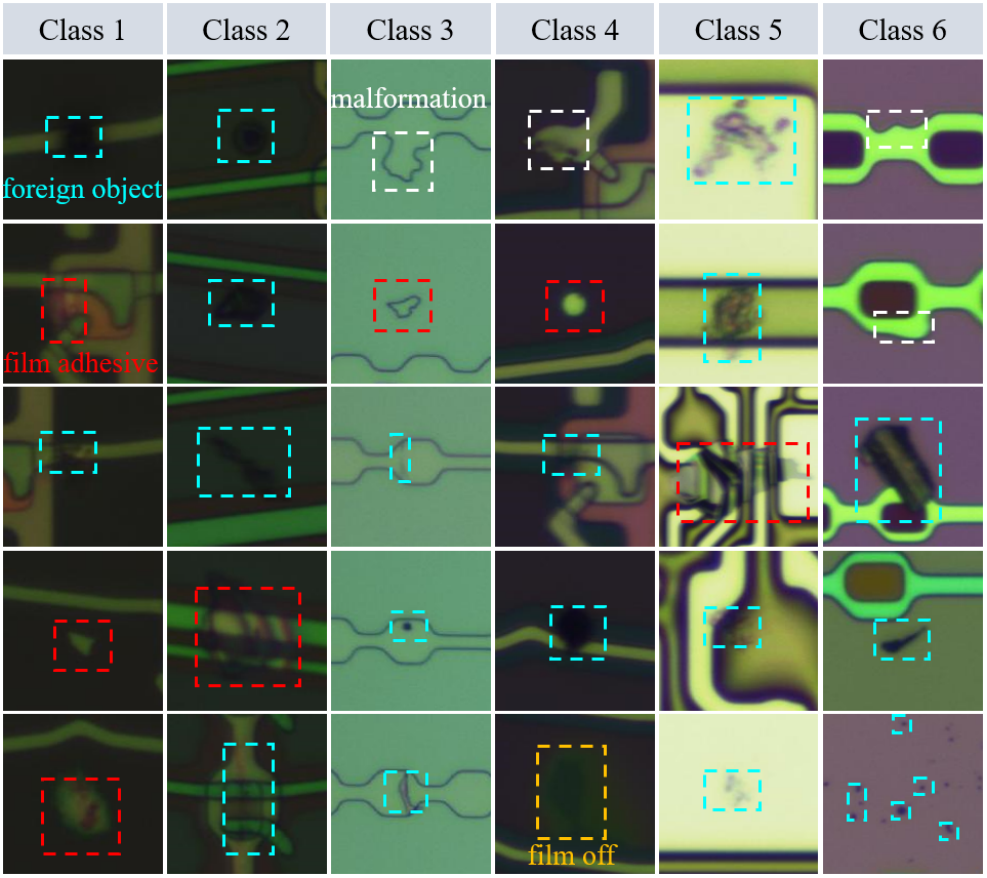

Figure 2: Différents défauts dans six classes de fonds de panneaux d'affichage.

Méthodes d'apprentissage Profond à la Pointe du Progrès

Les réseaux de neurones convolutifs (RNCs) sont des réseaux de neurones à action directe dotés d'une structure profonde, qui effectuent des opérations mathématiques de convolution sur l'image d'entrée au moyen de noyaux convolutifs (matrice de poids) de certaines tailles et de certains nombres.

Techniquement, le premier modèle RNC a été proposé par LeCun et al. en 1989 (LeCun et al., 1989) et amélioré en 1998 (LeCun et al., 1998), qui est appelé LeNet-5 visant à reconnaître les chiffres manuscrits. La Fig. 3 montre l'architecture de ce modèle RNC, qui contient une couche d'entrée, deux couches convolutionnelles, deux couches de mise en commun et trois couches entièrement connectées (la dernière étant la couche de sortie).

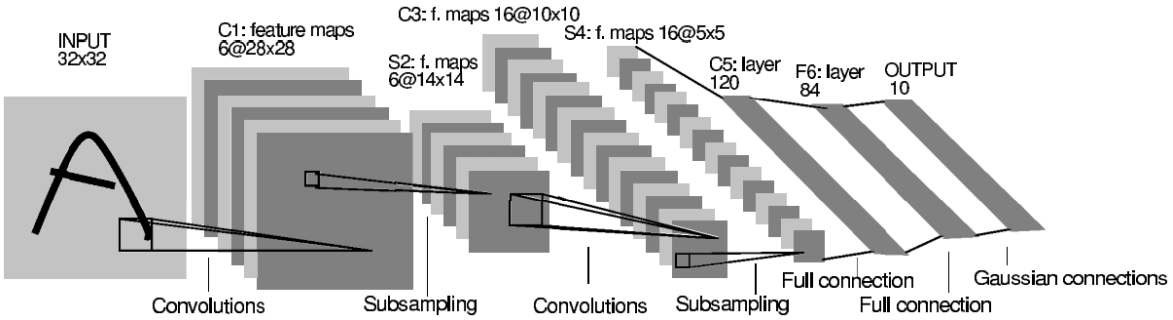

Figure 3: Architecture de LeNet-5 (LeCun et al., 1998). Chaque plan est une carte de caractéristiques.

En raison des bonnes performances de LeNet-5 dans la reconnaissance de chiffres manuscrits, il a reçu l'attention d'Alex Krizhevsky, dont le modèle AlexNet (Krizhevsky et al., 2012) avec un RNC plus profond proposé lors de l'événement " ImageNet Large Scale Visual Recognition Challenge " (alias " compétition ImageNet ") en 2012 a pu classer 1,2 million d'images haute résolution de 1000 catégories et a remporté le championnat. La Fig. 4 présente l'architecture du modèle AlexNet, qui contient une couche d'entrée, 5 couches convolutionnelles, trois couches de mise en commun et deux couches entièrement connectées.

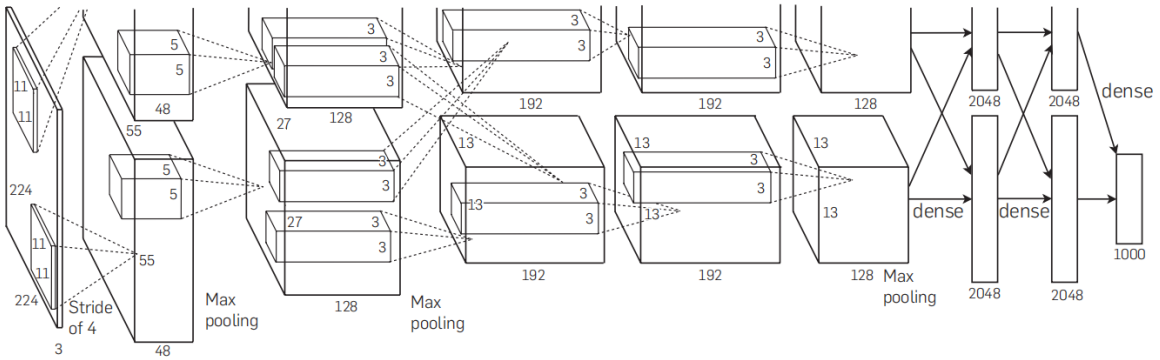

Figure 4: Une illustration de l'architecture d'AlexNet (Krizhevsky et al., 2012), montrant explicitement la délimitation des responsabilités entre les deux GPU.

En 2014, Simonyan et Zisserman ont proposé le modèle VGG (Simonyan and Zisserman, 2014a) avec une profondeur de 11 à 19 couches. La Fig. 5 montre l'architecture du modèle VGG16 comportant 16 couches. Ils ont remplacé le noyau de convolution de grande taille (11×11) d'AlexNet par de nombreux petits noyaux de taille 3×3, qui ont le même champ réceptif que le grand noyau et augmentent le nombre de couches de poids et d'unités non linéaires, améliorant ainsi les performances du réseau.

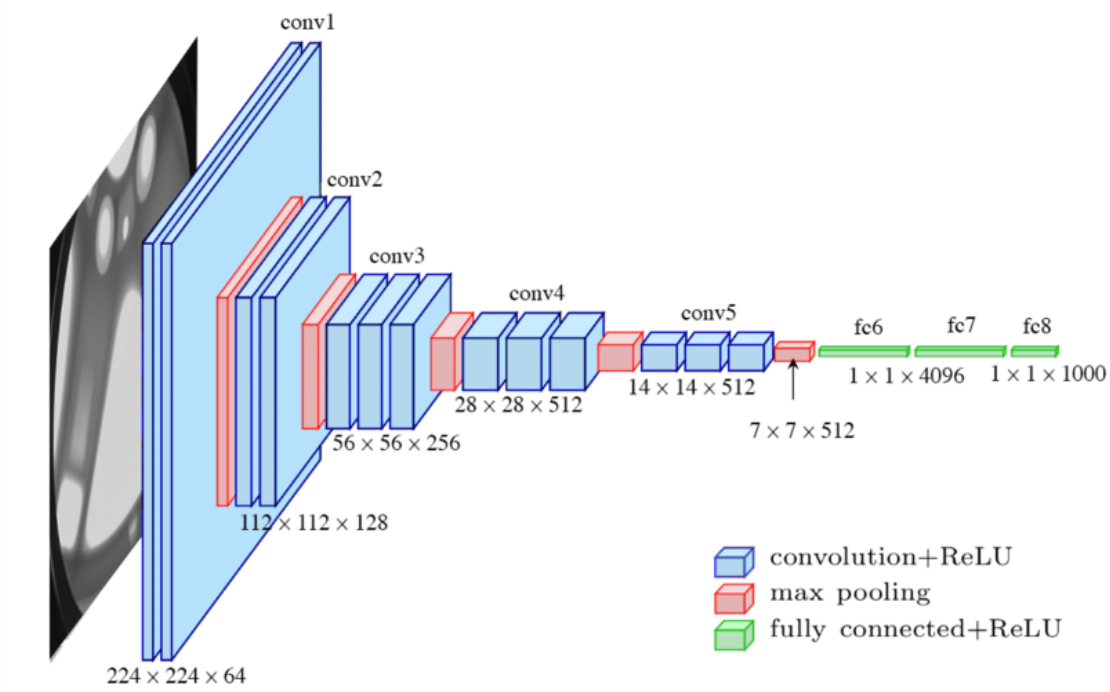

Figure 5: Architecture du VGG16 (Ferguson et al., 2017).

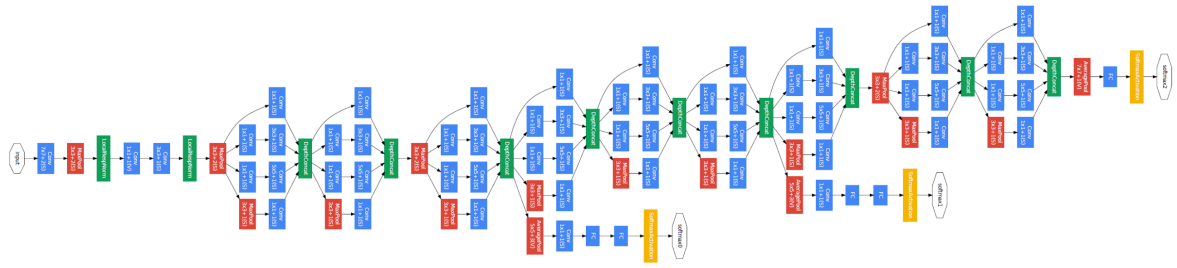

Figure 6: Architecture d'Inception-v1 (a.k.a. GoogLeNet) (Szegedy et al., 2015a).

Plus tard, en 2015, Szegedy et al. ont développé un GoogLeNet (Szegedy et al., 2015a) dont l'architecture est beaucoup plus profonde, et ils l'ont encore amélioré pour en faire Inception-v3 (Szegedy et al., 2016) en 2016 et Inception-v4 et Inception-ResNet (Szegedy et al., 2017) en 2017, respectivement. La Fig. 6 présente l'architecture de la première version, c'est-à-dire GoogLeNet.

Avec l'augmentation de la profondeur du réseau, son apprentissage devient plus complexe en raison du phénomène de décalage des covariables internes (c'est-à-dire que la distribution des entrées de chaque couche change avec les paramètres de la couche précédente). Ce phénomène rend nécessaire de fixer un taux d'apprentissage plus faible et de poser plus d'exigences pour l'initialisation des paramètres, ce qui alourdit inévitablement l'apprentissage, en particulier pour les réseaux présentant des non-linéarités saturantes. Par conséquent, Ioffe et Szegedy ont proposé la normalisation par lots (Ioffe and Szegedy, 2015) un mois après la publication de GoogLeNet. Elle résout le phénomène de décalage des covariables et permet d'entraîner le réseau

avec un taux d'apprentissage plus élevé et une initialisation grossière des paramètres. Depuis lors, les chercheurs ont utilisé la normalisation par lots comme un élément essentiel des RNC, généralement immédiatement après la couche convolutive et avant la fonction d'activation.

Il est à noter que tous les RNCs ci-dessus sont développés pour résoudre des problèmes de classification, c'est-à-dire qu'ils ne sortent que les probabilités correspondant aux catégories. Cependant, dans de nombreux cas, il est nécessaire d'obtenir la position (comme la reconnaissance de cible) ou le contour entier de l'objet cible (comme la segmentation sémantique), et c'est le problème de localisation ou de segmentation qui est un point chaud de la recherche actuelle. Le premier modèle RNC pour les tâches de segmentation qui a fait l'objet d'une grande attention et qui est considéré comme le pionnier est le FCN (c'est-à-dire le réseau entièrement convolutif), qui a été proposé (Long et al., 2015) par Long et al. en 2015. La Fig. 7 illustre l'architecture du modèle FCN. Il a remplacé toutes les couches entièrement connectées des RNC de classification précédents par des couches convolutives de mêmes dimensions, ce qui permet au RNC d'apprendre les caractéristiques des images d'entrée de grande taille et d'extraire des informations pertinentes pour la localisation et la limite de la cible. Il a ensuite ajouté des couches d'échantillonnage ascendant pour ramener la résolution de la sortie finale à celle de l'entrée du réseau par déconvolution, c'est-à-dire l'inverse de la convolution. La carte récupérée est le résultat indiquant la localisation ou la segmentation. En outre, le modèle FCN a utilisé une méthode appelée "skip" pour combiner les connaissances des couches profondes et peu profondes afin de réaliser une prédiction dense, ce qui a considérablement amélioré la précision de la segmentation.

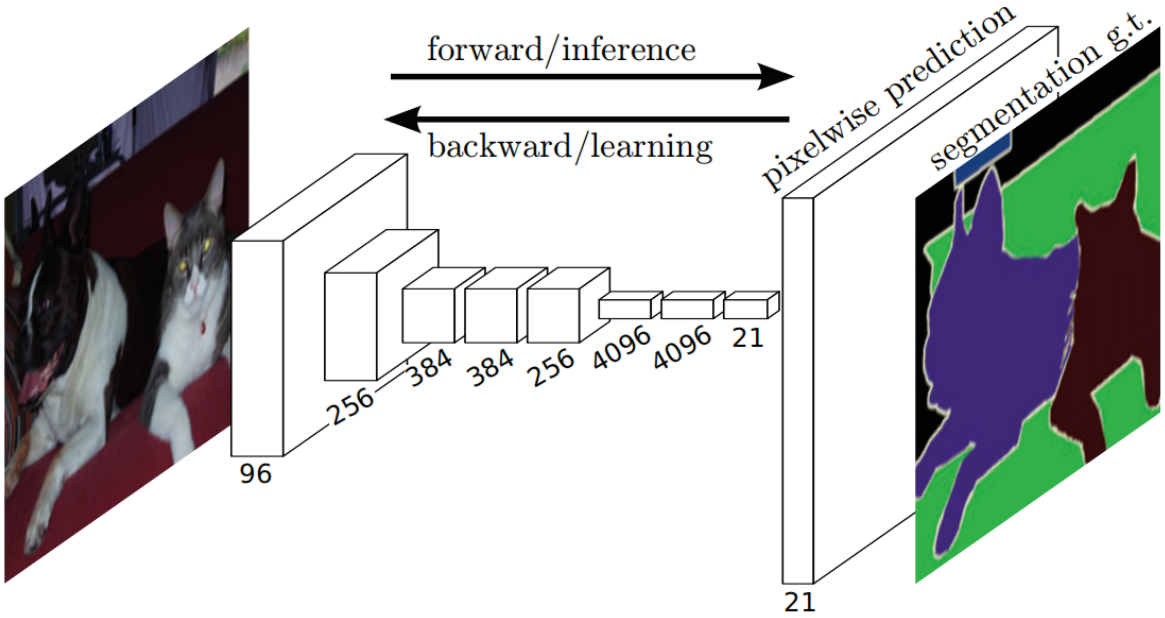

Figure 7: Architecture du FCN (Long et al., 2015).

L'apprentissage supervisé est la méthode d'apprentissage la plus couramment utilisée pour les modèles d'apprentissage profond, qui nécessite des ensembles de données d'entraînement avec des étiquettes complètes. En particulier, pour les

tâches de segmentation ou de localisation, l'étiquette au niveau du pixel indiquant l'emplacement de l'objet et les informations sur le contour (région cible) est nécessaire pour l'apprentissage supervisé. Cependant, dans de nombreux cas, l'obtention d'étiquettes solides au niveau du pixel est très coûteuse et prend beaucoup de temps, alors que des étiquettes relativement faibles au niveau de l'image sont beaucoup plus faciles à produire. Ce fait a conduit à la naissance et au développement florissant d'une autre méthode d'apprentissage, à savoir l'apprentissage faiblement supervisé. Le principe de l'apprentissage faiblement supervisé pour la segmentation ou la localisation (aussi appelé segmentation ou localisation faiblement supervisée) est qu'un modèle d'apprentissage profond est censé fournir des annotations par pixel sur la base de l'apprentissage avec seulement des étiquettes par image.

Résumé

Dans ce chapitre, nous décrivons les objets de recherche et la méthode d'apprentissage profond pertinents pour cette thèse:

- Tout d'abord, ce chapitre présente une introduction plus détaillée au diagnostic du cancer du sein, y compris les types de cancer du sein, les techniques d'imagerie pour le diagnostic des maladies du sein, et deux ensembles de données publiques pour l'étude des algorithmes des systèmes de diagnostic assisté par ordinateur utilisés pour aider les médecins ou les radiologues à prendre des décisions de classification plus précises et objectives pour les images histopathologiques cliniques.

- Ensuite, ce chapitre présente une introduction à la détection des défauts des panneaux d'affichage, y compris le processus de fabrication des panneaux d'affichage, les algorithmes des systèmes d'inspection optique automatisés pour la détection des défauts en ligne, et l'ensemble de données pour étudier les algorithmes des systèmes d'inspection optique automatisés qui sont appliqués pour soulager la pression de l'inspection visuelle manuelle par le personnel de l'atelier et pour réaliser une détection des défauts en ligne très efficace.

- Enfin, ce chapitre décrit deux techniques étroitement liées à cette thèse. L'une est celle des réseaux de neurones convolutifs en présentant leur origine et leur développement ainsi que deux algorithmes importants qui sous-tendent leur mise en œuvre. L'autre est l'apprentissage faiblement supervisé ainsi que sa définition et son développement.

Chapitre 3 Réseaux Neuronaux Convolutifs Configurables

Introduction

Au cours des deux dernières décennies, les méthodes d'apprentissage profond, en particulier les réseaux de neurones convolutifs (RNC), ont été le point le plus chaud de la recherche et ont été largement utilisées dans de nombreux domaines du traitement

d'images, tels que la détection de cibles (Han et al., 2022; Wang et al., 2022), la classification d'objets (Huang et al., 2017; Ma et al., 2018), et la segmentation sémantique (Lin et al., 2022; Zhou et al., 2022) en raison de leur apprentissage automatique des caractéristiques et de leurs capacités supérieures de représentation des caractéristiques. Plus précisément, les modèles VGG (Simonyan and Zisserman, 2014a), ResNet (He et al., 2016a), Inception (Szegedy et al., 2015a, 2016, 2017), et DenseNet (Huang et al., 2017) ont été développés pour réaliser la classification des images de nature (à partir de l'ensemble de données ImageNet) ; RCNN et ses séries (Girshick et al., 2014; He et al., 2017; Ren et al., 2015; He et al., 2017), YOLO et ses séries (Redmon et al., 2016; Redmon and Farhadi, 2017, 2018; Bochkovskiy et al., 2020; Wang et al., 2022), et BoxeR (Nguyen et al., 2022) ont été proposés pour la détection de cibles ; FCN (Long et al., 2015), U-Net et ses séries (Ronneberger et al., 2015; Alom et al., 2018; Schlemper et al., 2019; Thomas et al., 2020; Zunair and Hamza, 2021; Lin et al., 2022), SegNet (Badrinarayanan et al., 2017), et HSSN (Li et al., 2022) ont été développés pour la segmentation sémantique de différents objets.

Inspiré par les percées considérables et le succès des méthodes d'apprentissage profond dans les tâches de classification, de localisation et de segmentation en analyse d'images médicales (Ronneberger et al., 2015; Carneiro et al., 2017; Schlemper et al., 2019; Xu et al., 2019a; Ibtehaz and Rahman, 2020) et la détection de défauts (Zou et al., 2018; Lian et al., 2019; Liu et al., 2019; Dong et al., 2019; Song et al., 2020; Huang et al., 2021b), nous tentons de tirer parti de la performance supérieure des RNC pour atteindre nos objectifs. Cela conduit aux deux objectifs principaux de cette thèse mentionnés dans le chapitre 1. Tout d'abord, nous tentons de développer un modèle d'apprentissage profond pour réaliser une classification efficace et précise des images histopathologiques du cancer du sein tout en fournissant des explications visuelles (c'est-à-dire en introduisant l'explicabilité ou l'interprétabilité et la transparence au RNC qui a la nature d'une " boîte noire "), ce qui est d'une grande importance pour la sécurité, l'éthique, la confiance et la fiabilité du diagnostic clinique et pour le déploiement de systèmes de CAO basés sur l'apprentissage profond dans des contextes cliniques réels. Deuxièmement, nous tentons de développer un modèle d'apprentissage profond très efficace et précis pour les systèmes d'inspection optique automatisée (IOA) afin de résoudre le problème de la détection en temps réel des défauts des panneaux d'affichage sur les chaînes de montage des usines, qui joue un rôle essentiel dans l'amélioration du taux de rendement.

Néanmoins, l'architecture actuelle des réseaux d'apprentissage profond pour une tâche cible spécifique est généralement fixe et ne convient pas à d'autres tâches. Par exemple, un modèle RNC efficace pour une tâche de segmentation ou de localisation doit fournir des prédictions au niveau du pixel, ce qui nécessite que le modèle ait une structure d'encodeur-décodeur pour générer des résultats de bout en bout où le décodeur nécessite des processus de suréchantillonnage. Les tâches de classification exigent d'un modèle RNC qu'il extraie les informations les plus discriminantes de la région cible afin d'identifier avec précision les différentes catégories. Il est donc indispensable que le modèle RNC dispose d'un classificateur à la fin pour produire des probabilités de catégories, et les probabilités sont des prédictions au niveau de l'image. En outre, les différents objets d'investigation d'une même tâche présentent souvent de grandes variations, comme la segmentation sémantique des tumeurs mammaires (voir

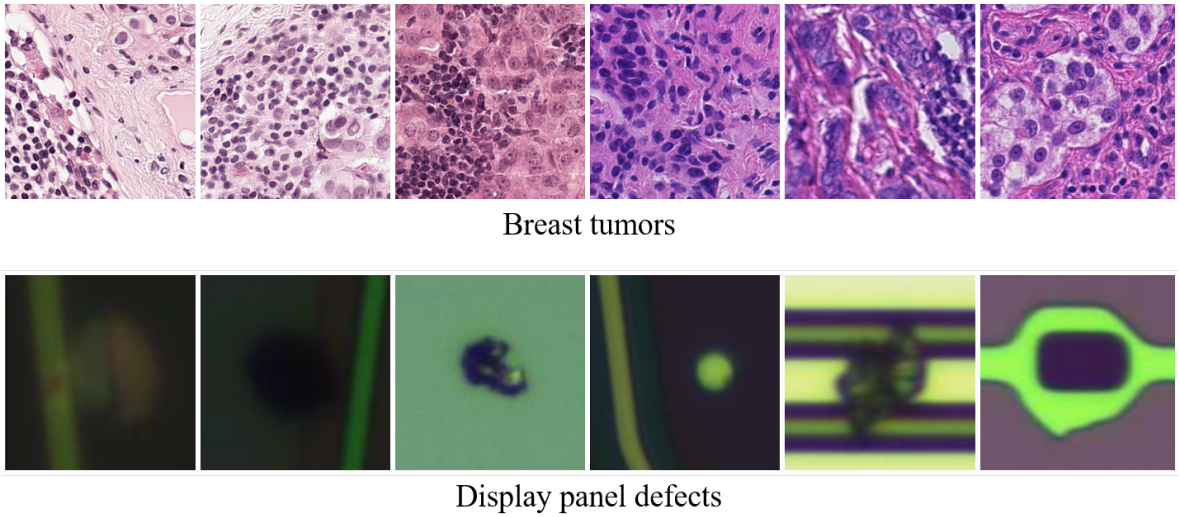

Figure 8: Exemples typiques d'images de tumeurs du sein (rangée du haut) et d'images de défauts de panneaux d'affichage (rangée du bas).

Fig. 8-haut) et les défauts des écrans (voir Fig. 8-bas). Comme il existe de nombreuses différences dans les textures de leurs images, les modèles d'apprentissage profond ont des exigences différentes pour l'extraction de caractéristiques de ces deux objets. En outre, le premier est plus favorable à la précision, tandis que le second est plus favorable à la vitesse de détection. Par conséquent, nous avons développé un réseau de neurones convolutif configurable (ConfigNet) capable de se transformer en différentes configurations pour s'adapter à de multiples tâches et objets, ce qui est compatible à la fois avec les classifications explicables d'images histopathologiques du cancer du sein et la détection en ligne des défauts des panneaux d'affichage.

Réseau Configurable Proposé

L'architecture globale de notre ConfigNet (c'est-à-dire le réseau configurable) est présentée à la Fig. 9. Elle est principalement composée de deux branches dorsales qui construisent deux configurations s'adaptant à différentes tâches (par exemple, la classification et la segmentation).

L'une d'elles est la configuration FEM-DMG-classifieur qui est appliquée pour réaliser une classification d'images explicables, où FEM (c'est-à-dire le module d'extraction de caractéristiques) est développé pour extraire les caractéristiques cibles, DMG (c'est-à-dire le générateur de cartes de décision) est développé pour générer des cartes de décision (c'est-à-dire des cartes de confiance des catégories) pour les explications, et le classifieur est développé pour prendre des décisions. Comme on peut l'observer plus en détail dans la Fig. 10, il n'est entraîné qu'avec la vérité de base au niveau de l'image (par exemple, normale ou tumorale) et est capable de fournir des probabilités de catégories (classification) et des prédictions au niveau du pixel (explication). Afin d'exploiter les performances supérieures des réseaux d'apprentissage profond existants, le FEM est configuré via la stratégie d'apprentissage par transfert, c'est-à-dire en utilisant les couches convolutionnelles peu profondes des modèles RNC

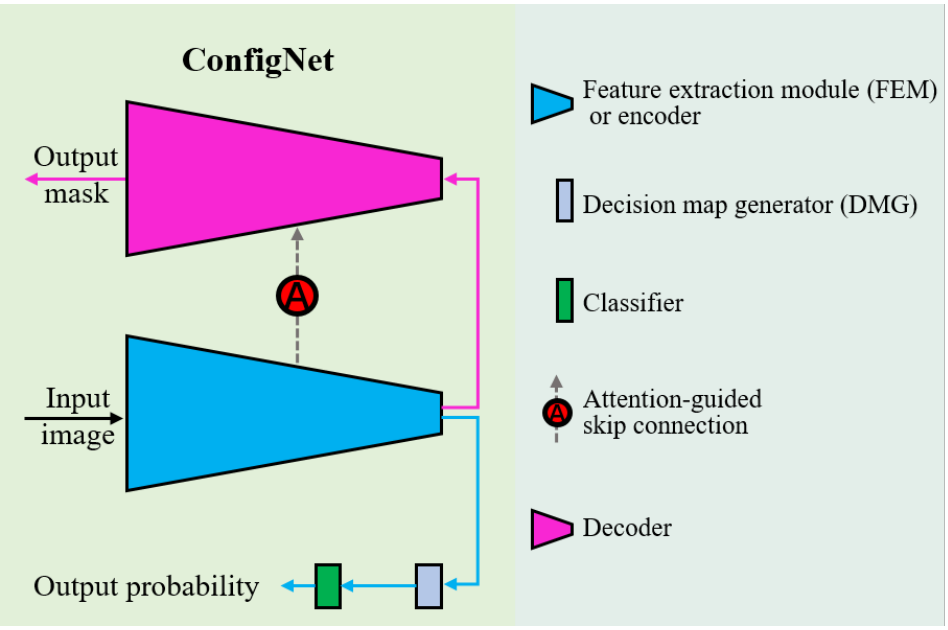

Figure 9: Illustration de ConfigNet

existants (tels que VGG (Simonyan and Zisserman, 2014a) et ResNet (He et al., 2016a)) comme notre FEM. Le DMG est un module qui fait face aux caractéristiques abstraites extraites de FEM et génère des cartes avec leurs canaux correspondant directement aux catégories, fournissant ainsi des preuves pour la classification suivante. Deux configurations de la DMG ont été développées dans cette thèse. L'une est la DMG de base avec une structure simple, et l'autre est la DMG améliorée avec des champs réceptifs de tailles multiples. Le classificateur est utilisé pour modéliser les cartes de décision du DMG en probabilités avec une correspondance biunivoque entre les canaux des cartes et les catégories. La mise en commun de la moyenne globale est un classificateur classique à cette fin. Nous avons proposé un classificateur de pooling moyen pondéré (WAP) pour la configuration FEM-DMG-classificateur de notre ConfigNet afin d'obtenir de meilleures performances.

L'autre configuration est celle de l'encodeur-décodeur qui est appliquée pour réaliser la segmentation et la localisation des cibles, où l'encodeur (alias FEM) est un module qui code l'image d'entrée en cartes avec des centaines et des milliers de canaux contenant différentes caractéristiques de l'image, tandis que le décodeur avec une connexion de saut guidée par l'attention doit transformer ces cartes de caractéristiques en masques (par exemple, un masque binaire) avec des régions (chacune contient des pixels de même valeur) indiquant les emplacements et les contours des cibles. La Fig. 11 présente plus en détail la configuration de l'encodeur-décodeur. Il est entraîné de bout en bout avec la vérité terrain au niveau du pixel. Comme pour la configuration du classificateur FEM-DMG, l'encodeur est construit sur la base de l'apprentissage par transfert et du réglage fin des RNC existants avec des couches convolutionnelles profondes ayant des performances remarquables dans l'extraction de caractéristiques. Le décodeur est construit en fonction de l'encodeur et possède une structure approximativement symétrique, qui peut être configurée pour favoriser différents objectifs. La connexion par saut guidée par l'attention (module de fusion des caractéristiques,

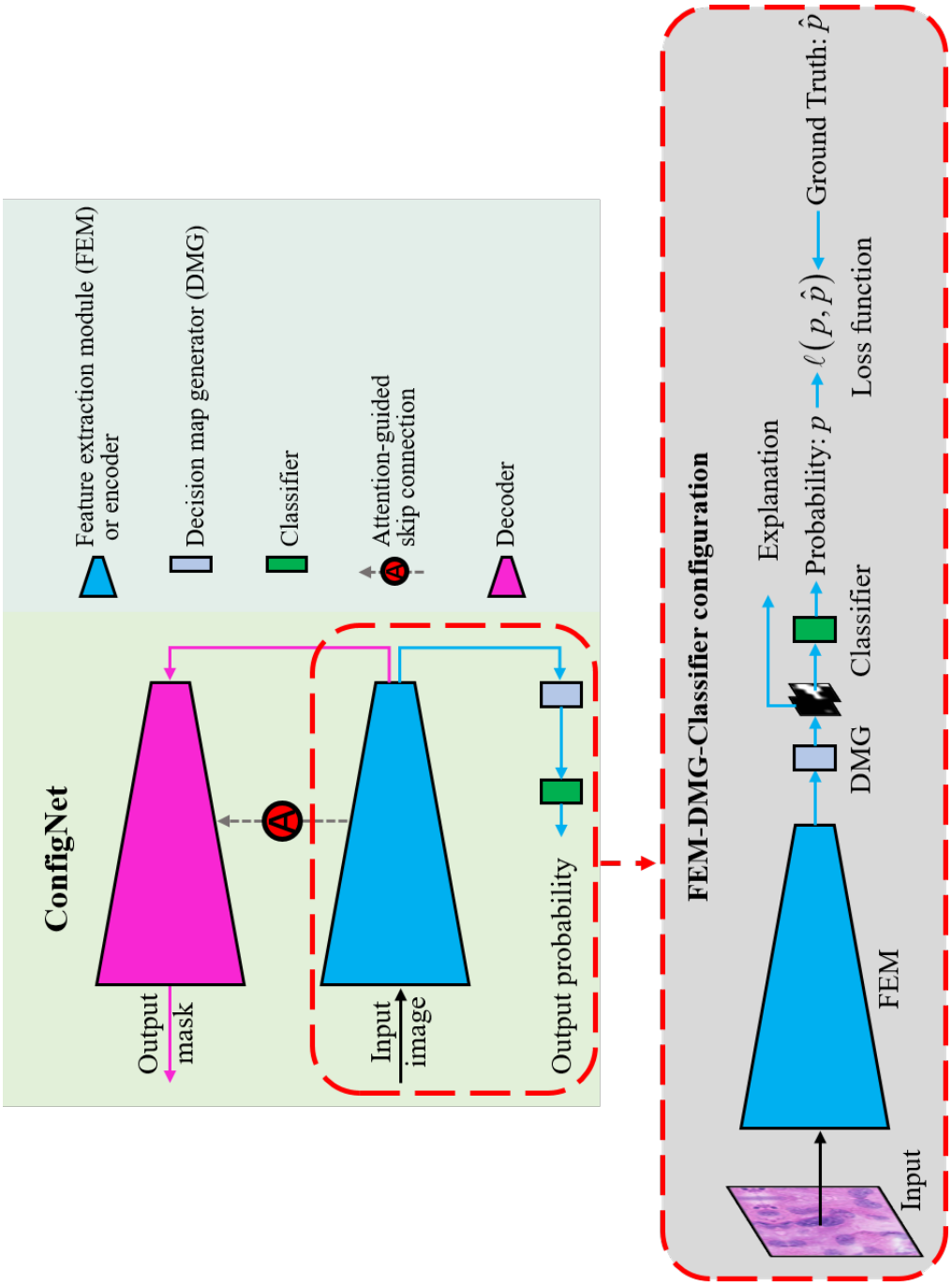

Figure 10: Illustration de la configuration du classificateur FEM-DMG.

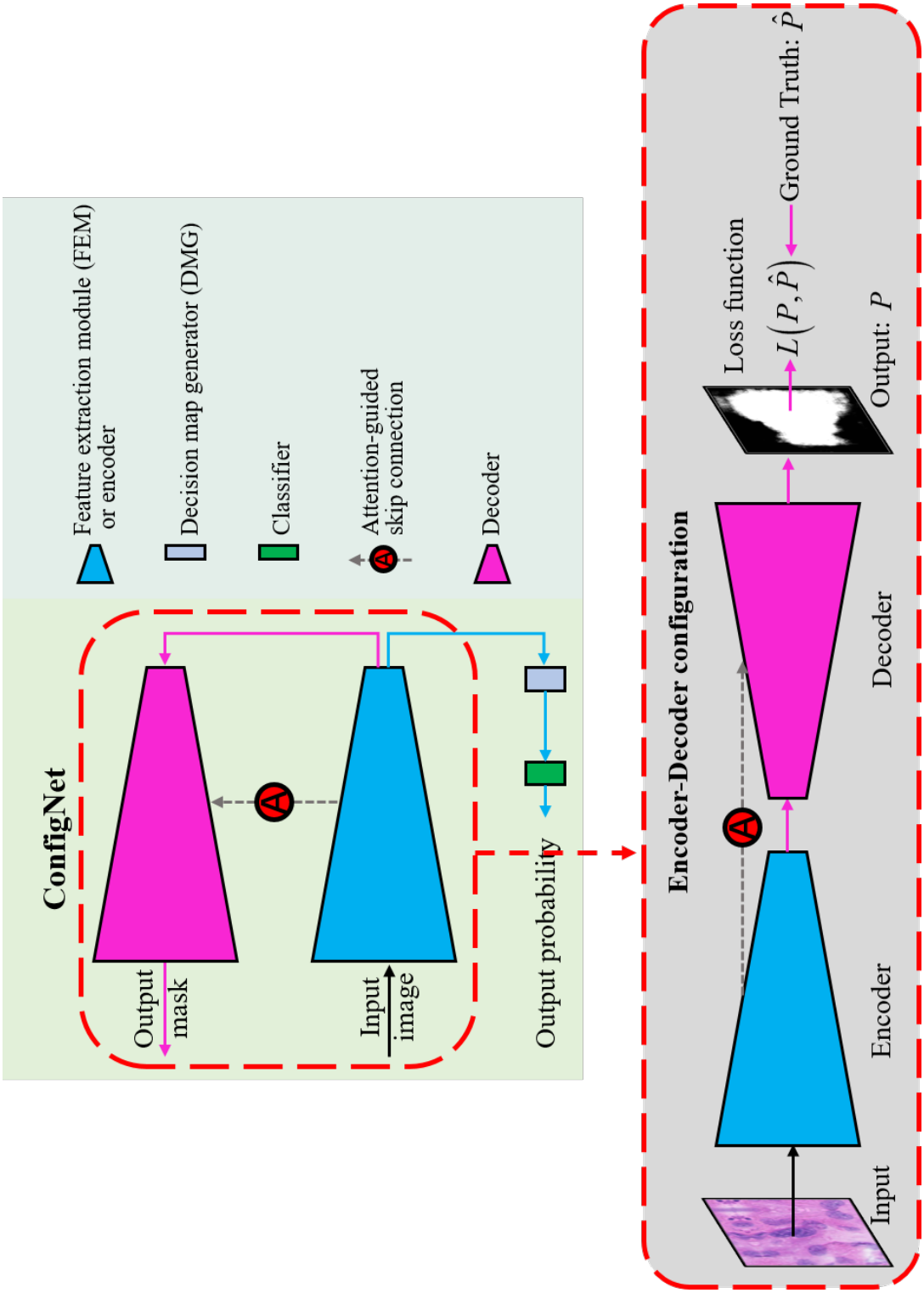

Figure 11: Illustration de la configuration de l'encodeur-décodeur.

abrégé FFM) entre l'encodeur et le décodeur (c'est-à-dire les couches superficielles et profondes) est développée pour encourager le réseau à se concentrer davantage sur les régions cibles tout en discréditant les réponses du fond. Nous avons proposé deux configurations du FFM pour la configuration encodeur-décodeur de notre ConfigNet. L'une est un module de fusion de caractéristiques par éléments (EFFM) favorisant l'efficacité, construit sur la porte d'attention de l'Attention U-Net (Schlemper et al., 2019) et utilisé pour la détection de défauts en ligne. Un autre module de fusion de caractéristiques guidées par l'attention spatiale et de canal (SCAFFM), privilégiant la précision, est conçu pour la segmentation d'images médicales.

Pour vérifier l'efficacité de notre ConfigNet dans de multiples tâches, telles que la classification explicable et la segmentation des tumeurs du cancer du sein, nous avons réalisé des expériences sur deux configurations, respectivement, sur deux ensembles de données du cancer du sein disponibles publiquement. La première est la configuration FEM-DMG-classifieur basée sur le VGG, le DMG de base et le classifieur WAP. La seconde est la configuration codeur-décodeur basée sur VGG et SCAFFM.

Résumé

Ce chapitre présente principalement la méthodologie clé développée pour cette thèse:

- Tout d'abord, nous donnons une description détaillée du réseau neuronal convolutif configurable (ConfigNet) proposé, y compris l'architecture globale du ConfigNet, la structure détaillée (y compris la stratégie de supervision) de la configuration du classificateur FEM-DMG développée pour la classification explicable d'images médicales, et la structure détaillée de la configuration de l'encodeur-décodeur développée pour la segmentation ou la localisation d'objets.

- Deuxièmement, nous menons des expériences approfondies pour la configuration du classificateur FEM-DMG (nommée MICNet) basée sur le classificateur VGG11, DMG de base et WAP sur deux ensembles de données de cancer du sein disponibles publiquement et nous la comparons à de nombreuses autres méthodes d'apprentissage profond pour évaluer la performance de classification de notre ConfigNet sur les images histopathologiques de cancer du sein. Les résultats de la comparaison quantitative sur le jeu de données BreakHis démontrent la performance de classification supérieure du MICNet, et les résultats visuels du MICNet sur le jeu de données basé sur le patch Camelyon16 vérifient son efficacité à fournir une explication raisonnable de la classification.

- Enfin, nous mettons en œuvre une expérience de comparaison de la configuration codeur-décodeur (appelée SCAFFNet) basée sur le SCAFFM développé et le décodeur " goulot d'étranglement " avec d'autres modèles de segmentation de pointe afin d'évaluer la performance de segmentation de la ConfigNet. Les résultats expérimentaux démontrent que notre SCAFFNet surpasse les autres modèles de segmentation dans plusieurs métriques d'évaluation, y compris mIoU, l'indice Dice du premier plan et de l'arrière-plan, la précision et le FPR. En outre, le SCAFFNet présente des performances de segmentation plus fines sur les tumeurs mammaires aux limites complexes.

Chapitre 4 ExplaCNet: Classification explicable d'images histopathologiques du cancer du sein basée sur un apprentissage faiblement supervisé

Introduction

Les progrès considérables réalisés dans le domaine de la classification explicable soulèvent de grands défis. Du fait qu'elle n'a pas accès à l'inférence directe du réseau, la méthode indépendante du modèle explique principalement pourquoi le réseau prend une certaine décision sans se soucier de sa justesse. Elle ne répond pas non plus à la contrainte selon laquelle les régions cibles indiquées par une carte d'explication visuelle cliniquement compréhensible par l'homme doivent mettre en évidence le tissu de la lésion de manière aussi complète que possible et présenter moins de faux positifs. La satisfaction de cette contrainte est la condition préalable à une carte explicative logique et cliniquement fiable. La méthode spécifique au modèle se heurte à un conflit crucial entre de meilleures performances de classification et une meilleure carte d'explication, bien qu'elle soit plus étroitement liée à la décision du réseau. Ce conflit est particulièrement important dans la classification clinique d'images histopathologiques où les différences interclasses entre les tissus tumoraux et normaux sont moindres et où les différences intraclasses entre les tumeurs ne sont pas négligeables.

Compte tenu de la nécessité de respecter la contrainte d'explication visuelle dans les contextes cliniques réels et de la grande importance de la connexion logique entre l'explication et la décision, nous avons proposé une configuration de notre ConfigNet favorisant l'explicabilité, à savoir un module d'extraction de caractéristiques (FEM)-générateur de cartes de décision (DMG)-classifieur. La configuration FEM-DMG-classifieur, appelée ExplaCNet, est une méthode explicable spécifique au modèle. Elle permet de résoudre le problème susmentionné et de réaliser la classification visuellement explicable des images histopathologiques du cancer du sein. Nous avons démontré la performance supérieure de notre méthode dans le diagnostic cliniquement pertinent des images histopathologiques du cancer du sein par des expériences complètes sur le jeu de données public Camelyon16 basé sur les patchs et sur le jeu de données BreakHis. De plus, nous avons rapporté l'étude d'ablation pour évaluer et vérifier l'efficacité des principaux composants utilisés dans notre ExplaCNet. Nos principales contributions sont les suivantes :

- Nous avons proposé un nouveau cadre explicable, ExplaCNet, basé sur WSL pour imiter la logique de diagnostic des experts humains en fournissant des cartes d'explication visuelle qui indiquent la localisation et les régions des tissus lésionnels. Il respecte les critères de transparence étroitement liés à la sécurité, à l'éthique et à la fiabilité du diagnostic clinique et apporte un soutien grand au déploiement de systèmes de CAO basés sur des modèles d'apprentissage profond dans des contextes cliniques réels.

- Nous avons adopté la stratégie d'apprentissage par instances multiples (MIL) pour améliorer la capacité de notre réseau à traiter des images de grande taille et nous avons développé un classificateur adaptatif de mise en commun des

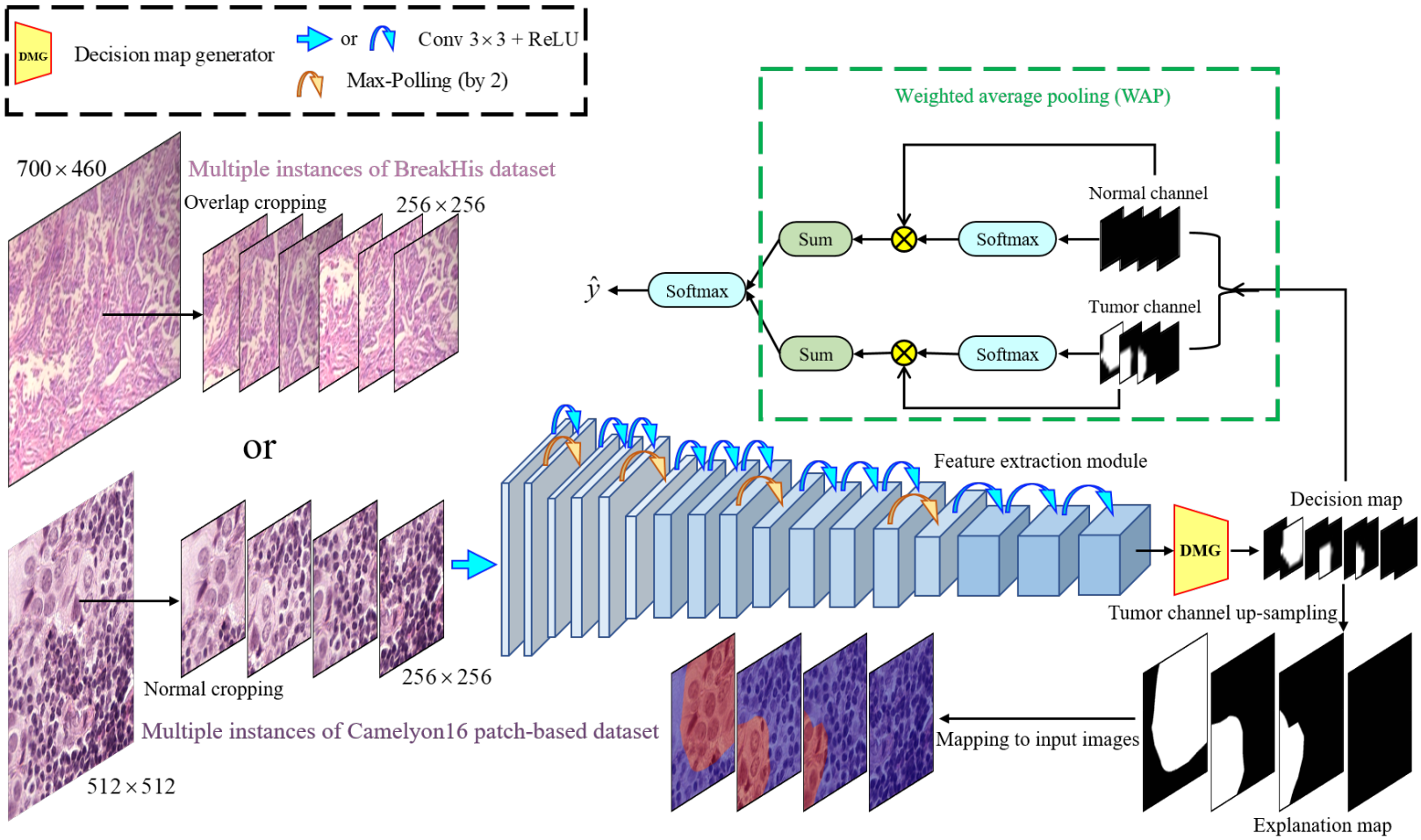

Figure 12: Une illustration de l'architecture ExplaCNet.

moyennes pondérées (WAP) pour fusionner les instances multiples et modéliser la probabilité finale du sac d'instances. La stratégie MIL améliore non seulement les performances de notre ExplaCNet en matière d'identification de tissus normaux et de classification d'images, mais elle ouvre également la voie au traitement d'images pathologiques de lames entières contenant des milliards de pixels. Le classificateur WAP adaptatif permet au réseau de reconnaître un plus grand nombre de régions lésionnelles, ce qui diminue les faux négatifs.

- Nous avons développé un DMG composé de filtres multi-échelles pour encourager le réseau à faire un compromis raisonnable entre les sensibilités aux tissus tumoraux et normaux, améliorant ainsi les performances du réseau dans l'identification des tissus des deux catégories et générant finalement une carte de décision avec des régions lésionnelles et normales séparées plus correctement pour la classification et l'explication.

Méthodologie

L'architecture de notre ExplaCNet, y compris le prétraitement des images d'entrée, est présentée à la Fig. 12. La méthode proposée comporte trois composants clés, à savoir, MIL, DMG et WAP.

Au cours de la procédure de diagnostic ainsi que de l'entraînement, les images d'entrée ont d'abord été découpées en un sac d'instances qui ont ensuite été introduites dans un FEM pour obtenir des cartes de caractéristiques contenant suffisamment d'informations contextuelles et discriminantes relatives à la cible. Ensuite, un DMG a été utilisé pour traiter les caractéristiques profondes codées et générer une carte de décision qui indique la présence de cibles et leur localisation. Enfin, la carte de décision résultante a été introduite dans un classificateur WAP suivi d'une fonction Softmax pour produire la probabilité du sac d'instances, c'est-à-dire la probabilité de classification de l'image d'entrée.

Notez que le FEM utilisé dans notre méthode a été construit sur l'architecture VGG16 (sans normalisation de lot en raison de notre petite taille de lot, c'est-à-dire 1) (Simonyan and Zisserman, 2014b) pré-entraîné avec le jeu de données ImageNet ; nous avons utilisé la première cinquième couche (sur un total de six couches, y compris la couche du classificateur) de ce modèle pour coder les images d'entrée. Nous avons décrit en détail le classificateur MIL et WAP dans le Chapitre 3 et nous donnerons une description détaillée du DMG proposé dans ce chapitre.

Conclusion

Nous avons proposé un nouveau cadre WSL appelé ExplaCNet pour réaliser la classification explicable des images histopathologiques du cancer du sein dans le diagnostic clinique. La stratégie MIL a été utilisée pour découper une entrée de grande taille en plusieurs instances afin d'encourager ExplaCNet à identifier davantage de tissus sains. Un classificateur WAP a été développé pour forcer ExplaCNet à reconnaître plus de régions tumorales. En outre, nous avons conçu un module DMG au-dessus d'un extracteur de caractéristiques profond afin d'optimiser l'attention du réseau sur les pixels

tumoraux et normaux et de générer des cartes explicatives fiables et compréhensibles par l'homme pour soutenir les résultats de la classification.

Des résultats expérimentaux approfondis sur le jeu de données Camelyon16 basé sur des patchs et sur le jeu de données BreakHis ont démontré que notre ExplaCNet surpasse les modèles explicables de pointe en matière d'explication visuelle (c'est-à-dire la fiabilité de la classification, qui est d'une importance capitale pour les méthodes d'apprentissage profond à déployer dans des contextes cliniques réels) tout en conservant une performance compétitive en matière de classification. Cette performance supérieure rend notre ExplaCNet plus adapté au diagnostic clinique. En outre, l'étude sur l'ablation a permis de vérifier l'efficacité de chaque composant clé de notre réseau.

Il convient de noter que notre cadre WSL peut être facilement adapté aux architectures de classification existantes afin d'assurer l'explicabilité et qu'il présente un grand potentiel pour la réalisation d'un diagnostic explicable sur l'ensemble de la surface. À l'avenir, il serait intéressant de mettre notre méthode en pratique et d'étudier le compromis optimal entre la génération de cartes explicatives plus compréhensibles par l'homme et l'obtention de résultats de classification plus précis dans le contexte clinique.

Chapitre 5 EFFNet: Réseau de Fusion de Caractéristiques par Eléments Pour la Détection des Défauts des Panneaux d'Affichage

Introduction

Figure 13: Défauts typiques difficiles à reconnaître sur les panneaux d'affichage. (a) Le faible défaut dans le fond clair. (b) Le défaut fort dans le fond de couleur sombre. (c) Le défaut de malformation avec une limite ambiguë. Les marques rouges indiquent les régions de défauts.

Bien que les modèles RNC aient fait des progrès considérables dans la détection des défauts, le problème de la détection des défauts en ligne et en temps réel n'a toujours pas été traité de manière exhaustive, notamment pour l'inspection de la production

de panneaux d'affichage. Tout d'abord, l'apprentissage profond est une technique axée sur les données et la performance d'un modèle RNC est fortement soumise aux données d'entraînement. Or, l'obtention d'un ensemble de données d'entraînement de panneaux d'affichage suffisamment nombreux et variés est à la fois chronophage et laborieux. D'autre part, la détection des défauts dans les panneaux d'affichage est confrontée à de nombreux défis non résolus : 1) le faible contraste entre le défaut et le fond environnant se présente sous deux formes : des défauts faibles sur un fond de couleur claire (Fig. 13(a)) et des défauts forts sur un fond de couleur sombre (Fig. 13(b)) ; 2) un bruit de fond complexe et diversifié sur les panneaux d'affichage ; 3) des défauts de malformation aux limites ambiguës : ces défauts sur les panneaux d'affichage ont une couleur identique à la zone sans défaut environnante, qui ne peut être distinguée par la différence d'intensité (Fig. 13(c)). Ces facteurs imposent sans aucun doute des exigences strictes pour la conception d'un modèle RNC rapide et précis.

Inspirés par le réseau U-Net d'attention (Schlemper et al., 2019), qui présente des performances impressionnantes dans les tâches de segmentation, même avec une faible quantité de données d'entraînement étiquetées, nous avons développé une configuration encodeur-décodeur de notre ConfigNet privilégiant l'efficacité, afin de surmonter les problèmes susmentionnés et de réaliser une détection en ligne des défauts des panneaux d'affichage avec une grande précision dans le monde réel. Cette configuration est appelée EFFNet, c'est-à-dire réseau de fusion de caractéristiques par éléments. Compte tenu de la difficulté de rassembler un ensemble de données suffisamment important sur les défauts des panneaux d'affichage, une stratégie d'apprentissage par transfert et une augmentation des données ont été utilisées dans notre méthode.

Les principales contributions de ce chapitre sont les suivantes :

- Un nouveau modèle EFFNet basé sur VGG16 et une architecture d'encodeur-décodeur a été proposé pour résoudre le problème de la détection en temps réel des défauts des panneaux d'affichage avec des arrière-plans à classes multiples.
- À notre connaissance, il s'agit de la première tentative d'adoption d'un modèle RNC intégré au mécanisme d'attention additive pour résoudre le problème de la détection en ligne des défauts des panneaux d'affichage après le processus dit d'array.
- Nous avons conçu une stratégie d'apprentissage par transfert et par réglage fin pour le module d'extraction de caractéristiques. Elle permet d'accélérer l'apprentissage du réseau et d'améliorer les performances de détection.
- Nous avons développé un module de fusion de caractéristiques basé sur l'addition par éléments de caractéristiques pyramidales appariées en taille. Il met en évidence les régions d'intérêt (ROI) sur le même canal des cartes de caractéristiques entre les caractéristiques peu profondes et profondes, ce qui réduit considérablement le temps de détection et améliore l'identification des pixels défectueux.

Méthodologie

L'architecture globale de notre modèle EFFNet, telle qu'elle est décrite à la figure 5.2, se compose d'un encodeur et d'un décodeur. Le codeur, c'est-à-dire le module

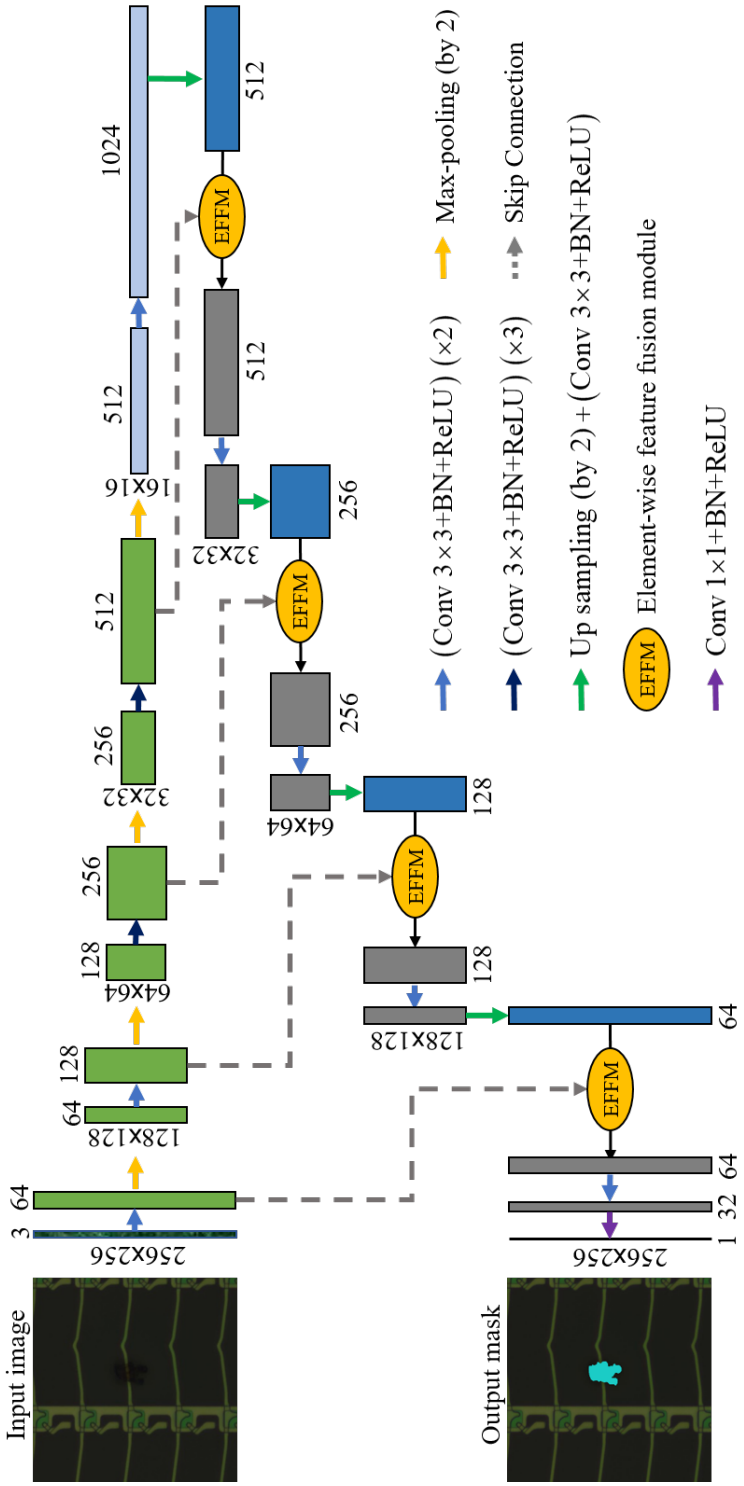

Figure 14: Une illustration du modèle EFFNet proposé.

d'extraction des défauts, extrait les caractéristiques des défauts d'un arrière-plan complexe couche par couche et forme une structure pyramidale de caractéristiques spatiales qui contient les informations sémantiques des défauts à différents niveaux. D'autre part, le décodeur intégrant des modules de fusion des caractéristiques par éléments intègre progressivement les informations sur les défauts à partir des caractéristiques codées. Après les processus ci-dessus, les résultats de la segmentation qui indiquent la localisation des défauts dans les panneaux d'affichage sont obtenus.

Conclusion

Nous avons proposé un nouveau modèle EFFNet pour la détection en ligne des défauts, qui revêt une grande importance pour le contrôle de la qualité et l'amélioration du taux de rendement des panneaux d'affichage. Les ConvBlocks VGG-16 entraînés par ImageNet modifiés et une stratégie de réglage fin ont été introduits pour le codeur afin d'extraire de manière extensive les caractéristiques complexes des défauts. En outre, un module de fusion des caractéristiques par éléments (EFFM) basé sur le mécanisme d'attention additive a été développé pour notre décodeur afin de fusionner les caractéristiques à plusieurs niveaux pour améliorer la précision de la détection tout en évitant une plus grande complexité de calcul. Les résultats expérimentaux montrent que notre méthode est plus performante que les méthodes de détection des défauts les plus récentes et que les autres techniques de segmentation. Elle présente également une bonne robustesse face au flou de mouvement et à un petit ensemble de données d'entraînement. De plus, notre modèle peut détecter les défauts à des vitesses acceptables pour la détection de défauts en temps réel.

Néanmoins, il reste encore quelques améliorations à réaliser pour mettre en pratique le réseau développé. Dans les travaux futurs, il serait intéressant de concevoir une stratégie d'amélioration des données plus efficace pour éviter l'échec de la méthode proposée dans des cas plus difficiles et d'optimiser le cadre pour obtenir une précision encore plus élevée et une vitesse plus rapide pour la détection de différents défauts dans les panneaux d'affichage.

Chapitre 6 Conclusions Générales et Perspectives

Conclusions

Dans cette thèse, nous avons présenté le travail original de développement de méthodes d'apprentissage profond efficaces pour la classification explicable d'images histopathologiques de cancer du sein et la détection de défauts en ligne de panneaux d'affichage. Afin de présenter notre travail de manière plus complète, nous avons donné d'abord une brève introduction à l'énoncé du problème dans le chapitre 1, qui mène à nos objectifs. En outre, les objets à étudier et le contexte connexe ont été présentés dans le chapitre 2, y compris le diagnostic du cancer du sein, la détection des défauts de l'écran et les méthodes d'apprentissage profond de pointe pertinentes. Les principales contributions des chapitres 3 à 5 sont énumérées ci-dessous:

- Dans le chapitre 3, nous avons proposé un réseau neuronal convolutif configurable (ConfigNet) capable de se transformer en différentes configurations qui s'adaptent à plusieurs tâches et objets. Le ConfigNet contient principalement deux configurations fonctionnelles définies comme la configuration module d'extraction de caractéristiques (FEM)-générateur de cartes de décision (DMG)-classificateur pour la classification d'images explicables et la configuration encodeur-décodeur pour la segmentation et la localisation d'objets. Les piliers du module d'extraction de caractéristiques et de l'encodeur de ces deux configurations sont construits par l'apprentissage par transfert de RNC existants avec des couches convolutives profondes. Des expériences approfondies de la configuration FEM-DMG-classifieur (MICNet) basée sur l'apprentissage d'instances multiples (MIL), le VGG11, un DMG de base et un classificateur de mise en commun des moyennes pondérées (WAP) démontrent que notre MICNet surpasse les autres modèles RNC dans la classification des images histopathologiques du cancer du sein et peut fournir une explication visuelle logique qui soutient la prédiction du réseau. De plus, les résultats expérimentaux de la configuration codeur-décodeur (SCAFFNet) basée sur le module de fusion de caractéristiques guidé par l'attention spatiale et de canal (SCAFFM) et le décodeur avec une structure de " goulot d'étranglement " sur le jeu de données basé sur le patch Camelyon16 indiquent que le SCAFFNet surpasse les modèles de segmentation de pointe dans la segmentation des tumeurs du sein, en particulier dans le cas difficile où les tumeurs du sein ont des limites complexes.

- Dans le chapitre 4, nous avons proposé un nouveau réseau basé sur l'apprentissage faiblement supervisé appelé ExplaCNet qui est construit sur la configuration du classificateur FEM-DMG de notre ConfigNet pour réaliser la classification explicable des images histopathologiques du cancer du sein. Nous avons adopté le MIL qui encourage le réseau à identifier plus de tissus normaux et le classificateur WAP qui force l'ExplaCNet à apprendre à reconnaître plus de tissus de lésions. En particulier, nous avons développé un DMG avec des filtres de convolution multi-échelles qui permettent un compromis raisonnable dans la capacité de l'ExplaCNet à identifier les tissus normaux et de lésion pour générer des cartes de décision raffinées pour la classification finale et l'explication visuelle. Les résultats expérimentaux sur le jeu de données Camelyon16 basé sur les patchs et sur le jeu de données BreakHis démontrent que notre ExplaCNet surpasse les méthodes explicatives de l'état de l'art en termes d'explication visuelle tout en conservant une performance de classification compétitive.

- Dans le chapitre 5, nous avons proposé un nouveau réseau de fusion de caractéristiques par éléments (EFFNet) basé sur la configuration encodeur-décodeur du ConfigNet pour réaliser une détection en temps réel de haute précision des défauts des panneaux d'affichage. La méthode a adopté une stratégie d'apprentissage par transfert et par réglage fin pour les couches d'extraction de caractéristiques et un décodeur avec une complexité de calcul relativement faible. En particulier, un module de fusion de caractéristiques basé sur l'addition par éléments de caractéristiques pyramidales a été proposé dans le cadre de la connexion skip pour améliorer l'efficacité et la précision de la détection. L'EFFNet

a été comparé aux techniques traditionnelles de détection des défauts et aux modèles de pointe basés sur le RNC. Des expériences approfondies ont démontré que le réseau EFFNet peut détecter avec précision des défauts aux textures complexes, aux limites ambiguës et au faible contraste. Il présente également une bonne robustesse au flou de mouvement. Il surpasse les méthodes les plus récentes en termes de mIoU, MPA et mesure F1. De plus, elle est capable de détecter des défauts à des vitesses allant jusqu'à 159 fps/s avec des images d'entrée de taille 256×256.

En résumé, nous avons proposé un réseau neuronal convolutif configurable appelé ConfigNet, qui peut être configuré en différentes configurations adaptées à différents tâches et objets, comme la classification explicable d'images histopathologiques du cancer du sein et la détection en ligne des défauts des panneaux d'affichage. Les méthodes (configurations) proposées présentent des performances supérieures dans la classification explicable d'images de cancer du sein, qui a un grand potentiel pour soutenir le déploiement de systèmes de CAO basés sur l'apprentissage profond dans des contextes cliniques réels, ainsi que dans la détection en ligne des défauts des panneaux d'affichage, qui est d'une grande importance pour améliorer le taux de rendement.

Perspectives

Compte tenu des limites constatées dans l'étude actuelle, les travaux futurs sont discutés en deux volets. Le premier concerne le diagnostic du cancer du sein et implique l'étude d'images d'autres modes que les images histopathologiques, la classification de sous-catégories de cancer du sein et l'amélioration de la configuration du classificateur FEM-DMG. L'autre est la détection en ligne des défauts des panneaux d'affichage, ce qui implique l'étude d'un plus grand nombre d'échantillons de défauts de panneaux d'affichage du monde réel et l'amélioration de la configuration de l'encodeur-décodeur.

Pour le diagnostic du cancer du sein, les travaux futurs comprennent:

- Étudier la classification explicable d'images de cancer du sein d'autres modes, y compris la mammographie, l'échographie, le CT, l'IRM et le PET-CT sur la base de la configuration du classificateur FEM-DMG de notre ConfigNet afin d'étudier de manière exhaustive le rôle et la contribution de ConfigNet à la détection et au diagnostic du cancer du sein, ce qui facilite davantage le déploiement de systèmes de CAO basés sur l'apprentissage profond dans des contextes cliniques.

- Pousser l'étude de la classification explicable du cancer du sein plus loin dans les sous-catégories pour compléter la capacité de ConfigNet dans le diagnostic de tous les types de catégories de cancer du sein, aidant ainsi de manière plus complète les experts cliniques à faire des propositions de traitement optimales. Cela nécessite des annotations plus détaillées, c'est-à-dire des étiquettes au niveau de l'image des sous-catégories et de leurs masques au niveau du pixel, qui nécessitent d'importantes charges de travail.

- La mise en pratique de l'ExplaCNet développé et l'étude du compromis optimal entre la génération de cartes explicatives plus compréhensibles par l'homme

et l'obtention de résultats de classification plus précis dans le cadre clinique contribueraient à la réalisation de son application clinique. En outre, l'idée du transformateur (Vaswani et al., 2017) pourrait être utile pour étendre la configuration du classificateur FEM-DMG du ConfigNet afin de réaliser davantage de tâches d'analyse du cancer du sein.

Pour la détection en ligne des défauts des panneaux d'affichage, les travaux futurs incluent:

- Obtenir des ensembles de données plus diversifiés et plus complexes recueillis sur la chaîne de production de l'usine et étudier les performances de la configuration codeur-décodeur de notre ConfigNet sur ces ensembles de données. Cela permettrait d'améliorer ses performances et de renforcer sa fiabilité et sa stabilité. En outre, des études supplémentaires sur l'épine dorsale de la configuration codeur-décodeur basée sur des modèles plus légers tels que SqueezeNet (Iandola et al., 2016) pourraient être utiles pour son intégration dans les systèmes IOA d'utilisation terminale.

Chapter 1

General Introduction

1.1 Problem Statement and Objectives

The main objectives of this thesis are two-fold: One is to develop intrinsically explainable deep learning models for computer-aided diagnosis (CAD) systems to solve the problem of histopathological breast cancer image classification and provide trustworthy support for the deployment of deep learning-based CAD systems in real-world clinical settings. Another is to develop high-efficient and accurate deep learning models for automated optical inspection (AOI) systems to achieve online defect detection of display panels on factory assembly lines.

Breast cancer includes invasive and non-invasive, according to its site. Invasive breast cancer (Harris et al., 2016) occurs once ill cells from within the lobules or milk ducts split out into the proximity of breast tissue. It consists of many different subtypes, such as Infiltrating lobular carcinoma (ILC) (Yedjou et al., 2022), Infiltrating ductal carcinoma (IDC) (Zhou et al., 2021), Medullary carcinoma (Mateo et al., 2017), Mucinous carcinoma (Komenaka et al., 2004), Tubular carcinoma (Huang et al., 2021a), Inflammatory breast cancer (Hu et al., 2021), Paget disease of the breast (Sakorafas et al., 2001), Phyllodes tumor (PARK et al., 2021), and Triple-negative breast cancer (Li et al., 2019b). Non-invasive breast cancer (West et al., 2017), on the other hand, does not spread out from the lobules or ducts where it is situated. It mainly has two sub-types, i.e., ductal carcinoma in situ (DCIS) (Weedon-Fekjær et al., 2021) and lobular carcinoma in situ (LCIS) (Masannat et al., 2013) where "in situ" represents "in place." The complexity of its sub-type, its most prevalent occurrence, and women's negligence in breast self-examination and clinical examination make it the most deadly cancer for women cancer patients in the world, with a ratio of 1/4 among all cancer cases and 1/6 cancer deaths (685,000 deaths in 2020) (Akram et al., 2017; Sung et al., 2021).

The foundation of breast cancer regulations to improve treatment outcomes and survival lies in its early diagnosis, according to the World Health Organization (WHO). Thanks to the development of medical imaging technology, the regulations can recently be achieved in many approaches satisfying diverse requirements. Mammography (Gøtzsche and Jørgensen, 2013) is a standard mass screening technique that can image bone, soft tissue, and blood vessels all at the same time. Ultrasound (Ozmen et al., 2015) is a cost-effective and radiation-free technique that identifies cysts and solid masses, and it has been recommended as a supplement to evaluate lumps found in mammography.

Magnetic resonance imaging (MRI) (Kuhl, 2019) produces detailed images at different cross-sections through strong magnetic fields and computer-generated radio waves, which makes it able to detect small details of soft tissues that cannot be detected by mammography. Computed Tomography (CT) (Boone et al., 2006) scans a series of X-ray images at different angles and then uses tomographic reconstruction algorithms to produce cross-sectional images (slices) of the bones, blood vessels and soft tissues. Positron emission tomography (PET) (Vercher-Conejero et al., 2015) visualizes and measures changes in metabolic processes and other physiological activities (such as blood flow, regional chemical composition, and absorption) by injecting a peripheral vein with radioactive substances (radionuclides). Although the above medical imaging techniques are non-invasive and have specific advantages, they are either suffering from radiation risks, expensive costs, or low sensitivities (Onega et al., 2016; Hooley et al., 2013; Roganovic et al., 2015). Biopsy (Soo et al., 2019) is the only definitive way and "gold standard" to diagnose breast cancer, which removes a piece of tissue or a sample of cells from the body so it can be examined under a microscope. However, the biopsy diagnosis procedure is highly time-consuming and requires well-experienced human experts that need years of training. Fortunately, deep learning-based computer-aided diagnosis (CAD) systems (Qiu et al., 2017; Cong et al., 2020; Meng et al., 2021) that emerged in recent years have the potential to effectively alleviate most of the workloads of pathologists, thereby greatly ameliorating this problem.

Currently, deep learning methods have made great strides and achieved impressive results in object recognition and image processing (Chen et al., 2017; Voulodimos et al., 2018; Tian et al., 2020; Esteva et al., 2021; Zheng et al., 2021; Dong et al., 2022). Convolutional neural networks (CNNs), as the most influential products, are widely used for image classification and target localization due to their unique ability to learn complex features, such as VGG (Simonyan and Zisserman, 2014a), ResNet (He et al., 2016a), WILDCAT (Durand et al., 2017), and ACoL Zhang et al. (2018a). It has inspired many researchers (Spanhol et al., 2016a; Bayramoglu et al., 2016a; Sun et al., 2021; Song et al., 2017; Sudharshan et al., 2019; Kumar et al., 2020; Gour et al., 2020; Schirris et al., 2022; Xu et al., 2019a) to apply this method to investigate the classification of histopathological breast cancer images (i.e., the key procedure and ultimate goal of breast biopsy). Although deep learning methods have made tremendous progress in this area and even performed at par with human experts, most of the studies have ignored logical or rational explanations (i.e., failed to provide explainability) for the network decision. Introducing explainability into deep learning models is an essential way to unlock their "black box" nature that hinders their transparency, which is indispensable for the safety, ethics, human understanding, and reliability of clinical diagnosis. Lacking explainability has prevented the deployment of deep learning-based CAD systems in real-world clinical settings, which leads to the first goal of this thesis and indicates its great significance.

Similarly, as the display component of almost all modern electronic devices (such as smartwatches, cell phones, laptops, car center controls, and LCD TVs), the thin-film-transistor liquid-crystal display (TFT-LCD) and organic light-emitting diode (OLED) are in considerable demand worldwide. However, the manufacturing of both LCD and OLED is tedious and complicated. Take LCD as an example. Its production process is mainly divided into the front section array process, the middle section panel assembly

(Cell), and the last section module assembly (Module). Among these operations, the front section array process includes cleaning, film formation, photoresist coating, exposure, development, etching, and stripping. Although the manufacturing process is generally carried out in a clean room, various defects such as dust, dirt, short circuit, and breakage still inevitably occur on the product surface due to some technical reasons, which leads to defective products. The occurrence of these defective products seriously affects the yield of LCD and OLED, thus directly affecting profitability. Therefore, effective monitoring and inspection of this fabrication process during manufacturing and timely repair or elimination of defective products before packaging are essential to reduce waste of resources, reduce time cost, and improve yield rate.

Three types of panel defects detection methods have been adopted currently: human visual inspection, electrical inspection, and automated optical inspection (AOI). Human visual inspection is the traditional and primitive defect detection method, which is limited in two aspects, i.e., low detection accuracy in defects observed by the naked eye and inefficient detection of defects viewed by microscopic sampling. Meanwhile, it is subjective, uncertain, and easy to misjudge, and the storage and query of inspection data are inconvenient. This inspection method can not meet the needs of mass and rapid production. Electrical detection methods, such as probe scanning, photoelectric coupling, and conductive circuit detection, can only be used for functional defects caused by electrical factors and are not suitable for detection in the manufacturing process. AOI (He and Sun, 2015; Yuan et al., 2015; Wang et al., 2018; Tsai and Hung, 2005; Tsai et al., 2007) based on computer vision is the fastest-growing and most widely used method for surface defects detection due to its non-contact nature, high automation, and scalability. However, traditional computer vision techniques, such as Otsu (He and Sun, 2015), Canny edge detector (Wang et al., 2018), Fourier transform (Tsai and Hung, 2005; Tsai et al., 2007), Gaussian and Gabor filters (Tong et al., 2016), optical flow (Tsai et al., 2011), and support vector machine (SVM) (Chu et al., 2017), recognize defects primarily based on handcrafted features and rules that are limited in feature representation. They can hardly satisfy the requirement for industrial online defect detection of display panels having defects with low contrast, intricate and diversiform background noise, and ambiguous boundaries. Whereas, the currently developed deep learning methods (Liu et al., 2020a; Hu and Wang, 2020; Zou et al., 2018; Song et al., 2020; Lee et al., 2022; Li and Li, 2021; Chang et al., 2022; Yao and Li, 2022; Li and Wang, 2022) with superior feature extraction and learning abilities and considerable successes in object recognition bring a solution to this issue, which inspired the second goal of this thesis.

In general, the network architecture of a deep learning model is fixed for a specific target task (e.g., segmentation, localization, or classification) and object of investigation (e.g., histopathological breast cancer image classification or display panel defect detection). For example, for segmentation or localization tasks, where the goal is to obtain pixel-level results (i.e., binary masks), it is required that the network model be able to extract the approximate outline and localization information of the target. Current mainstream network architectures of this aspect mainly consist of encoder-decoder structures that enable end-to-end output (i.e., the input resolution is equal to the output resolution), such as U-Net (Ronneberger et al., 2015) and its series (Alom et al., 2018; Schlemper et al., 2019; Thomas et al., 2020; Zunair and Hamza, 2021; Lin et al.,

2022), SegNet (Badrinarayanan et al., 2017), DeepCrack (Zou et al., 2018), and EDRNet (Song et al., 2020). For classification tasks, the network model needs to recognize the category of the target in the image (e.g., benign or malignant), which requires it to be able to extract the most discriminative information relevant to the target region. Current classification networks are primarily composed of a feature extraction module (i.e., full convolutional layers with certain depth and width) and a classification module (e.g., fully connected layers or linear layers), such as VGG (Simonyan and Zisserman, 2014a), ResNet (He et al., 2016a), Inception (Szegedy et al., 2015a, 2016, 2017), DenseNet (Huang et al., 2017), and ShuffleNet (Zhang et al., 2018b; Ma et al., 2018). Thus, most of the existing deep learning models are task-oriented (i.e., lacking generality), which is detrimental to the investigation and implementation of our work. Based on a comprehensive study of previously reported works, we proposed a deep learning framework named configurable network (ConfigNet) that can be configured adaptively for different tasks and objects. It can be applied for achieving both explainable classification of histopathological breast cancer images and online defect detection of display panels.

1.2 Main Contributions

The main contributions of this thesis are detailed as follows:

- Configurable Convolutional Neural Networks (Chapter 3).

  We proposed a configurable convolutional neural network (ConfigNet) capable of transforming into different configurations that adapt to multiple tasks and objects. The ConfigNet mainly has two functional configurations defined as FEM-DMG-classifier configuration for explainable image classification and encoder-decoder configuration for object segmentation and localization. The FEM-DMG-classifier configuration is composed of a feature extraction module (FEM), a decision map generator (DMG), and a classifier. The backbones of the FEM and encoder of these two configurations are both constructed through the transfer learning of existing CNNs with deep convolutional layers. For the FEM-DMG-classifier configuration (MICNet) based on VGG11, a basic DMG, and a weighted average pooling (WAP) classifier, extensive experiments on BreakHis and Camelyon16 patch-based datasets demonstrate that it outperforms other CNN models in the classification of histopathological breast cancer images and can provide a logical visual explanation that supports the network prediction. For the encoder-decoder configuration (SCAFFNet) based on spatial and channel attention-guided feature fusion module (SCAFFM) and decoder with a "bottleneck" structure, experimental results on the Camelyon16 patch-based dataset show that it outperforms the state-of-the-art segmentation models in breast tumor segmentation, especially in the challenging case where the breast tumors have complex boundaries.

- ExplaCNet: Explainable classification of histopathological breast cancer images based on weakly-supervised learning (Chapter 4).

  We proposed a novel weakly-supervised learning-based network, ExplaCNet, to achieve the explainable classification of histopathological breast cancer images.

We used multiple instance learning (MIL) to encourage the ExplaCNet to identify more normal tissues and an adaptive weighted average pooling (WAP) classifier that fuses multiple instances into a bag probability to force the ExplaCNet to learn to recognize more lesion tissues. In particular, we developed a decision map generator (DMG) with multi-scale filters that allow a reasonable compromise in the ability of ExplaCNet to identify normal and lesion tissues to generate refined decision maps for final classification and visual explanation. Experimental results on Camelyon16 patch-based dataset and BreakHis dataset demonstrated that our ExplaCNet outperforms state-of-the-art explainable methods in terms of visual explanation while remaining a competitive classification performance.

- EFFNet: Element-wise feature fusion network for defect detection of display panels (Chapter 5).

  We developd a novel element-wise feature fusion network (EFFNet) to achieve high-accuracy real-time defect detection of display panels. The method adopted a transfer learning and fine-tuning strategy for feature extraction layers and a decoder with relatively less computational complexity. Particularly, a feature fusion module based on element-wise addition of pyramid features was proposed in skip connection to improve detection efficiency and accuracy. Our method was compared with both traditional defect detection techniques and state-of-the-art CNN-based models. Additionally, the effects of training dataset size, motion blur noise, and different backgrounds on the performance of the proposed method are investigated. Extensive experiments demonstrate that the developed network can accurately detect defects with complex textures, ambiguous boundaries and low contrast. It also has good robustness against motion blur noise. It outperforms state-of-the-art methods in terms of mIoU, MPA, and F1-Measure. Moreover, it is able to detect defects at speeds of up to 159 fps/s with input images of size 256×256.

1.3 Organization of Thesis

The thesis manuscript is organized as follows:

In Chapter 2, entitled "Objects and Background," we provide a detailed introduction to breast cancer diagnosis, including types of breast cancer, breast imaging techniques, and two public datasets of histopathological images for studying computer-aided diagnostic system algorithms. We also give an introduction to display panel defect detection, including display panel manufacturing, online defect inspection, and the dataset of display panel defect images for studying automated optical inspection system algorithms. Furthermore, we describe two deep learning methods closely related to this thesis. One is convolutional neural networks, and the other is weakly-supervised learning.

In Chapter 3, entitled "Configurable Convolutional Neural Networks," we proposed a configurable convolutional neural network named ConfigNet which can be configured adaptively for different tasks and objects to achieve the explainable classification of histopathological breast cancer images and online display panel defect

detection. The outline of the developed ConfigNet is first introduced. Secondly, the performance of its FEM-DMG-classifier configuration (named MICNet) for achieving the classification of histopathological breast cancer images while providing logical visual explanations is investigated. The whole architecture of our MICNet is described in detail, including MIL strategy, FEM, and weighted average pooling (WAP) classifier, as well as how it generates explanations. The performance of our MICNet was evaluated on two publicly available datasets. Finally, the encoder-decoder configuration (named SCAFFNet) based on spatial and channel attention-guided feature fusion module (SCAFFM) is compared with other state-of-the-art segmentation models on the Camelyon16 patch-based dataset to evaluate the segmentation performance, where the SCAFFNet is introduced in detail.

In Chapter 4, entitled "ExplaCNet: Explainable Classification of Histopathological Breast Cancer Images Based on Weakly-Supervised Learning," we developed an explainability-favored FEM-DMG-classifier configuration (named ExplaCNet) of our ConfigNet to address the conflict between providing better human-understandable explanation maps and producing more accurate classification decisions and achieve the clinically explainable classification of histopathological breast cancer images. The newly designed DMG with multi-scale filters, allowing a reasonable trade-off in the conflict and enhancing the explainability while keeping the classification performance, was detailed. Furthermore, an ablation study was performed to demonstrate the effectiveness of the key components of our ExplaCNet. Our ExplaCNet was evaluated in terms of both explanation and classification on two breast cancer datasets, along with many state-of-the-art deep learning methods.

In Chapter 5, entitled "EFFNet: Element-Wise Feature Fusion Network for Defect Detection of Display Panels," We proposed an efficiency-favored encoder-decoder configuration (named EFFNet, i.e., element-wise feature fusion network) of our ConfigNet for online defect detection of display panels. The EFFNet contains a defect extraction module, a feature decoder, and an element-wise feature fusion module (EFFM), which were introduced in detail. Ablation studies about the EFFM and different transfer learning strategies were implemented. Extensive experiments, including comparisons with non-deep learning and deep learning methods and discussion on the effects of training dataset size, motion blur noise, and backgrounds, were conducted to demonstrate the superiority and robustness of our EFFNet.

In Chapter 6, entitled "General Conclusions and Perspectives," we summarize the main contents of this thesis, including our major contributions, general conclusions, and future perspectives.

Chapter 2

Objects and Background

2.1 Breast Cancer Diagnosis 40

2.1.1 Types of Breast Cancer 41

2.1.2 Breast Imaging 45

2.1.3 Datasets for Investigating the CAD System of Breast Cancer Diagnosis 52

2.2 Display Panel Defect Detection 54

2.2.1 Display Panel Manufacturing 54

2.2.2 Online Defect Inspection Algorithms 57

2.2.3 Dataset for Investigating the AOI System of Online Defect Detection of Display Panels 60

2.3 State-of-the-Art Deep Learning Methods 61

2.3.1 Convolutional Neural Networks (CNNs) 61

2.3.2 Weakly-Supervised Learning 71

2.4 Summary 76

2.1 Breast Cancer Diagnosis

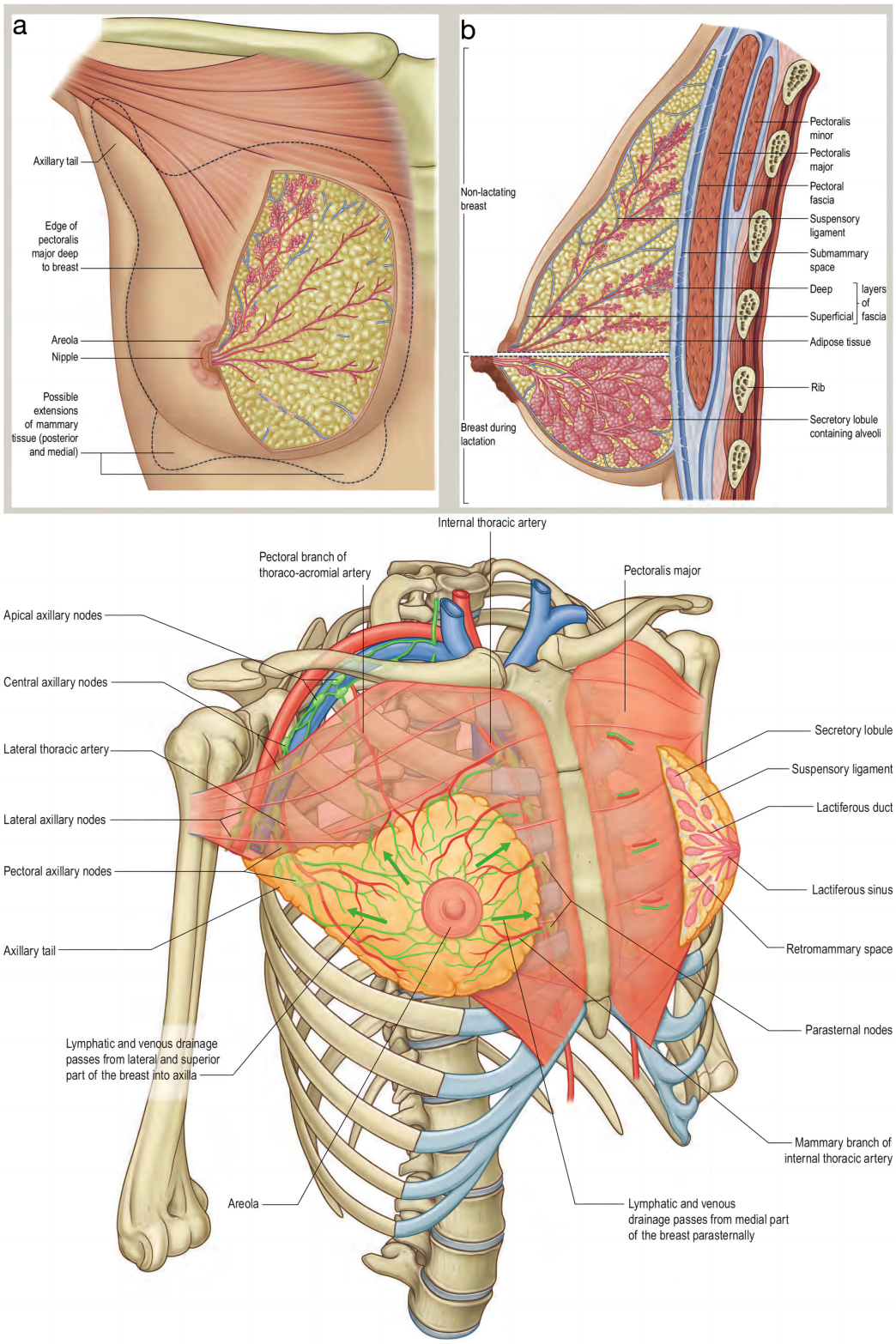

Figure 2.1: Breast anatomy (Bazira et al., 2021).

The breast (as shown in Fig. 2.1) is a tubuloalveolar gland surrounded by fatty adipose tissue (including fibrous connective tissue and ligaments) containing a network of nerves, blood vessels, lymph vessels, and lymph nodes (Stingl et al., 2005; Delotte et al., 2009; Fahad Ullah, 2019). In particular, female breasts normally consist of 12–20 lobes that are subdivided into many smaller lobules, and they are connected via milk ducts (Tanis et al., 2001). The mammary gland structure generally contains a stratified epithelium, consisting of myoepithelium and epithelium, bounded by a basement membrane and anchored in a template of vascular, lymphatic and stromal cells (Stingl et al., 2006). Breast cancer occurs for the following reasons: 1) The human immune system is unable to identify damage to deoxyribonucleic acid (DNA), errors in genes such as P53, BRCA1, and BRCA2, and inherited alterations and resolve them, leading to tumorigenic changes in myoepithelial cells or epithelial cells (or stem cells with the ability to develop into myoepithelial cells or epithelial cells), which in turn lead to cancer; 2) the RAS/MEK/ERK and PI3K/AKT pathways protect normal cells from cellular suicide. When the genes involved in encoding these protective pathways are mutated, cells fail to commit suicide when no longer needed, leading to the development of cancers associated with estrogen exposure (Akram et al., 2017). In cancer cells, telomerase reverses chromosome shortening and allows for extensive cellular replication (Hanahan and Weinberg, 2000). While tumor cells generally obtain their nutrient and oxygen supply through angiogenesis, cancer cells break this boundary and can enter the bloodstream, lymphatic tissue, and other tissues of the body to produce secondary tumors (Jain, 2005; Gupta and Massagué, 2006).

2.1.1 Types of Breast Cancer

Infiltrating Lobular Carcinoma

The breast is essentially a complex ductal vesicle-like gland, and the lobule is the basic unit of breast structure and function. It consists of 10-15 terminal enlarged acinus, the acinus ducts that are continuous with the acinus, and the terminal ducts that are continuous with the acinus ducts. In clinical practice, the distal end of the terminal duct, the acinus duct, and the anatomical area of the acinus are the main sites of breast cancer and various hyperplastic breast diseases. They are called the terminal duct lobe unit (TDLU). It is usually called invasive (or infiltrating) lobular carcinoma (ILC) (Dixon et al., 1982) when the cancer cells break through the basement membrane of the terminal milk ducts or acinus and invade the interstitial components outside the lobules. The progression of ILC is from atypical lobular hyperplasia to carcinoma in situ and then becomes invasive carcinoma. The ILC is arranged in a target ring around the ducts and grows infiltratively, forming stellate foci and preserving the normal structure of the ducts, and spreading individually in the interstitial fibers.

Infiltrating Ductal Carcinoma

Invasive (or infiltrating) ductal carcinoma (Silverstein et al., 1994) is a very heterogeneous group of tumors and one of the most common types of invasive breast cancer (40% to 80% of cases). It originates in the milk ducts of the breast and extends into the duct wall, invading the adipose tissue of the breast, as well as other parts of the body.

The cells are usually arranged in different structures, such as cords, trabeculae, masses, glandular ducts, and solid sheets. Some of them can have obvious central necrosis. Its cells show a variety of morphologies and are relatively large and adherent. The nuclear grade ranges from regular to markedly pleomorphic and has distinct nucleoli. There are different interstitial components such as fibroblasts, collagen fibers, elastic fibers, and lymphoplasmacytic, as well as necrosis and calcification.

Medullary Carcinoma

Medullary carcinoma (Jacquemier et al., 2005) is a type of breast cancer with poorly differentiated cells, a large patchy distribution, a lack of glandular structures and interstitium, and obvious lymphoplasmacytic infiltration. It has large cancer nests, little interstitium and a soft texture. Under the microscope, the cut surface is grayish-white, translucent, and resembles lymphoid tissue or brain marrow. The cancer cells are arranged in large pieces and cords (usually exceeding 4 layers of cells in width). It has swollen, extruded margins and a lack of infiltration into the surrounding breast and adipose tissue. More than 75% of cancer cells are syncytial cells with abundant cytoplasm. The tumor lacks adenoidal findings, and the tumor cells are poorly defined. The nuclei are vacuolated, with marked pleomorphism and heterogeneity, and some may show atypical tumor giant cells and squamous cell metaplasia. There is a large number of dense lymphocyte and plasma cell infiltration inside and outside the cancer foci. Extensive necrosis and hemorrhage were common in the cancerous tissue.

Mucinous Carcinoma

Mucinous carcinoma (Weigelt et al., 2008) (a.k.a. colloid carcinoma) is an uncommon breast cancer with nests of cells floating in lakes of mucin separated by delicate fibrous septae involving capillary blood vessels. Its large cell clusters have an occasional tubular arrangement and variable shapes and sizes. Classic mucinous carcinoma normally has a low nuclear atypia, which may prevail along with mitoses in rare cases Tan et al. (2008). A cribriform or micropapillary intraepithelial component is hardly seen. Mucinous carcinoma of type A that has larger quantities of extracellular mucin is a classic non-endocrine variety. It also has pure and mixed variants, and invasive carcinoma of no particular type is the most common admixture. The one composed of more than 90% mucinous carcinoma is a pure tumor. The in situ component may have a micropapillary, papillary or cribriform pattern and may appear predominant luminal mucin production.

Tubular Carcinoma

Tubular carcinoma (Cooper et al., 1978; Fernández-Aguilar et al., 2005) is a particular kind of invasive breast carcinoma with open tubules made up of one layer of epithelial cells that encloses a clear lumen. These features should comprise more than 90% of the tumor. The tubules are arranged haphazardly with rounded and angulated shapes or generally an admixture of an oval. The cells are scanty mitotic figures, inconspicuous nucleoli, small to moderate in size and regular with little nuclear pleomorphism.

Tubules do not have myoepithelial cells, but some may have a basement membrane with an incomplete surrounding layer. The cellular desmoplastic stroma is a secondary but essential feature commonly accompanied by tubular structures. The occurrence of tubular carcinoma is relevant to low-grade ductal carcinoma in situ, tubular neoplasia, and flat epithelial atypia. The proportion of tubular structures necessary for tubular carcinoma diagnosis is not fixed yet, but 90% purity is recommended. Tumors are supposed to be regarded as mixed type once having 50% to 90% tubules admixed with another morphology.

Inflammatory Breast Cancer

Inflammatory breast cancer (IBC) (Yeh et al., 2013; Joglekar-Javadekar et al., 2017) is a breast tumor with dimples and (or) wide ridges caused by cancer cells blocking the lymphatic vessels over the breast. It is uncommon to see tough inflammatory breast cancer, which is considerably fast-growing. The diagnosis of IBC requires an inter-professional approach (called clinicopathological diagnosis) which needs a core needle biopsy of the breast to make the initial diagnosis of invasive carcinoma. Tumor cells invading the dermal lymphatic is classic IBC seen on a breast biopsy. The local and metastatic disease occurs due to the same malignant cells from tumor emboli.

Paget Disease of the Breast

Paget disease of the breast (PDB) (Van der Putte et al., 1995) has a histologic hallmark of Paget cells presence, which are large malignant intraepithelial adenocarcinoma cells with variable sizes and presenting singly or in the form of small groups inside the nipple epidermis. Paget cells are conjectured to derive from glandular stem cells or clear cells of the nipple epithelium (epidermal Toker cells). The cells can be signet-ring, ovoid, or round forms, normally mucin positive. The cytoplasm may include periodic acid-Schiff (PAS)–positive, diastase-resistant granules that indicates neutral mucopolysaccharides presence. The cells have microscopic features of glandular cells that are pale to clear vacuolated cytoplasm. nuclei are usually high-grade with prominent nucleoli.

Phyllodes Tumor

Phyllodes tumors (Zhang and Kleer, 2016) are protruding, firm, and well-circumscribed masses. Its surface is tan or pink to grey and may exhibit as mucoid and fleshy. Large lesions have characteristic whorls with curved fissures, similar to leaf buds, but smaller ones may have a homogeneous appearance. Phyllodes tumors exhibit an enhanced growth pattern in the lumen with leaf-like protrusions into a varying degree of dilated and elongated ductal lumen. The epithelial component contains myoepithelial and luminal epithelial cells stretched into arc-like clefts surmounting stromal fronds. The stroma In benign phyllodes tumors is usually more cellular than in fibroadenomas. Malignant phyllode tumors exhibit a combination of marked nuclear pleomorphism of stromal cells, stromal overgrowth, stromal cells increase, mitosis increase, and border infiltration.

Triple-Negative Breast Cancer

Triple-negative breast cancer (TNBC) (Turner and Reis-Filho, 2006) refers to a group of breast cancers with negative expression of estrogen receptor, progesterone receptor, and human epidermal growth factor receptor. It is a subtype of breast cancer with unique biological and clinical features. Based on the molecular characteristics of triple-negative breast cancer, it can be categorized into six gene expression subtypes. Studies have shown an important association between gene expression subtypes of triple-negative breast cancer and specific mutation subtypes. Most triple-negative breast cancers are non-specific types of highly invasive ductal carcinomas characterized by high nuclear grade, high mitotic index, interstitial lymphocytic infiltration, central necrosis, and compression of adjacent tissues.

Lobular Carcinoma in Situ

Lobular carcinoma in situ (LCIS) (Chuba et al., 2005) has the terminal ducts or acinus within the lobules that are solidly expanded and filled with uniform and consistent tumor cells. The tumor cells are small and uniform in size and poorly adherent. The nuclei are round or ovoid form with uniform chromatin and inconspicuous nucleoli. The LCIS includes several subtypes: pleomorphic, florid adenosis, hyaline, and myxoid cell type. The most important is the pleomorphic subtype. In polymorphic LCIS, the tumor cells are poorly adherent. The nuclei are significantly enlarged, have prominent polymorphism and obvious nucleolus and nuclear division, and sometimes show acne-like necrosis or calcification. Atypical lobular hyperplasia (ALH) and LCIS have morphologic similarities, but the degree of involvement of the terminal ductal lobular unit (TDLU) differs. LCIS is diagnosed when more than 50% of the TDLU is filled with diagnostic cells and dilated, while ALH is diagnosed when it is less than 50%.

Ductal Carcinoma in Situ

Ductal carcinoma in situ (DCIS) (Burstein et al., 2004), a.k.a. intraductal carcinoma, is a non-invasive cancer. DCIS is classified into 3 grades, i.e., low-grade, intermediate-grade, and high-grade. High-grade DCIS often consists of larger pleomorphic cells with distinct nucleoli and common nuclear division. Acne-like necrosis with a large amount of necrotic debris is often seen in the lumen, but intraluminal necrosis is not necessary for the diagnosis of high-grade DCIS. Low-grade DCIS consists of small monomorphic cells with nuclei in round form and uniform size, uniform chromatin, inconspicuous nucleoli, and rare nuclear division. The tumor cells are arranged in a rigid hitchhiking bridge, micropapillary, sieve, or solid shape. The structural expression of intermediate-grade DCIS is diverse, and the cellular heterogeneity is between high-grade and low-grade DCIS.

2.1.2 Breast Imaging

Mammography

Mammography is an imaging technique that uses a low dose of X-rays (approximately 30 kVp) to examine the human breast. Normally, cancerous lumps and calcium deposits appear brighter on mammograms, thus enabling diagnostic and screening purposes.

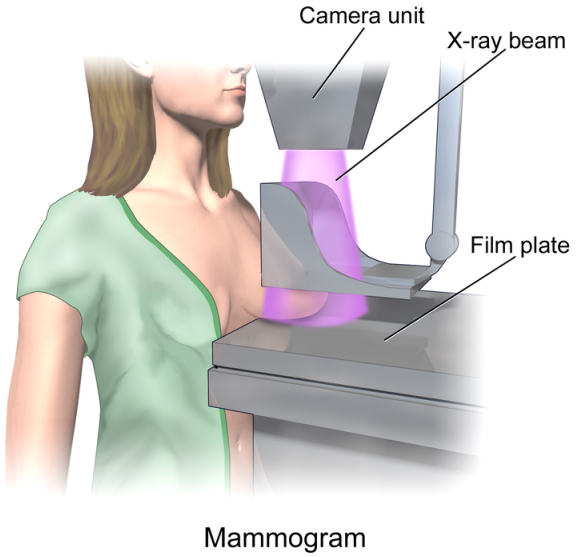

Figure 2.2: Illustration of a mammogram (Wikimedia, 2021).

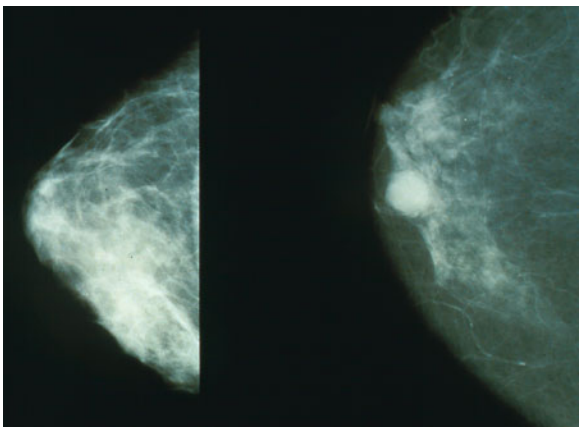

Figure 2.3: Normal (left) versus cancerous (right) mammography images (Wikimedia, 2019).

During an examination procedure (as shown in Fig. 2.2), the breast is placed and held still on a flat support plate and compressed with a parallel plate (a.k.a. a paddle) that evens the thickness of breast tissue, thus allowing low-energy X-rays to penetrate. Then, an X-ray machine produces a small burst of X-rays passing through the breast to a detector on the other side. The X-ray image (i.e., mammogram) can be captured either on a photographic film plate or a computer (digital image) that

receives electronic signals from a transmitting device. Screening mammography used for annual examination of breast sickness needs to take images of both head-to-foot (craniocaudal, CC) view and angled side-view (mediolateral oblique, MLO). Diagnostic mammography used for patients with breast symptoms may need more views of regions of concern, such as geometrically magnified and spot-compressed views.

In mammography films (as shown in Fig. 2.3), areas with different density distributions represent different tissues, including normal tissues, noncancerous masses of benign tumors, fibroadenomas, and complex cysts. Low-density tissue, such as fat, appears translucent (i.e., dark gray near the black background), while dense tissue, such as connective and glandular tissue or tumor, appears whiter on the gray background. A radiologist will determine the possibility of malignancy (i.e., cancer) by observing the shape, size, and contrast of abnormal regions and the appearance of the edges. Some very bright spots of tiny calcium fragments, called microcalcifications, are also under observation because they may be a sign of a specific type of cancer.

Mammography allows the examination of patient lesions with high definition and contrast, especially for ductal carcinoma in situ (DCIS) and calcifications (Sree et al., 2011). It can be used to observe the extent of tumor infiltration and detect the presence or absence of lymph node metastases in patients. It is mainly used for mass screening of breast disease (Jafari et al., 2018), to improve the treatment of early disease (Heywang-Köbrunner et al., 2011), and to reduce the mortality rate of breast cancer patients to some extent (Van Schoor et al., 2011). However, it also has many limitations. Mammography must be subjected to a certain amount of ionizing radiation, experiences a low sensitivity (decreases with increasing tissue density) and specificity, has a relatively high rate of false positives (Hofvind et al., 2012; Brodersen and Siersma, 2013) and false negatives (Hoff et al., 2012), and is prone to produce overdiagnosis (Kalager et al., 2012; Falk et al., 2013).

Ultrasound

Ultrasound (a.k.a. ultrasonography) is a non-invasive imaging technique that uses radiation-free high-frequency sound waves (7.5-15 MHz (Kossoff, 2000)) and corresponding echoes to visualize the inside of organs of concern. It enables the display of some changes in the breast, such as fluid-filled cysts and blood flow to the breast area, which is unlikely to be observed on a mammogram.

During the ultrasound procedure (as shown in Fig. 2.4), a transducer (a handheld, wand-like instrument) is adopted to move over the skin on and around the breasts. The gel is used between the skin and transducer to eliminate the air so as to avoid acoustic impedance and reflection and allow a clear image. The transducer first sends a beam of sound waves to the body, which then reflects back into the transducer by the boundaries between tissues in the beam path (such as between tissue and fluid or tissue and bone). Once these echoes come back to the transducer, an electrical signal is generated and sent to the ultrasound scanner. The scanner then calculates the distances between the transducer and the tissue boundaries according to the sound speed and return time of each echo. Finally, the two-dimensional images of the breast tissues are produced based on these distances.

Ultrasound imaging is the breast examination tool of choice for preoperative local-

ization of masses and can be used to achieve fast identification of cysts (as shown in Fig. 2.5) and solid masses. It, in a way, improves cancer detection rates in those at high risk for breast cancer. The location to be viewed can be specified throughout the imaging procedure, which makes it suitable for recording the localized lesion excision of the biopsy specimen. However, breast ultrasonography is unable to detect many tumors due to the very similar acoustic properties of healthy and cancerous tissues. It is therefore applied as an adjunct to mammography and clinical examination rather than as a stand-alone modality (Hooley et al., 2013; Ozmen et al., 2015). When ultrasonography is used as a supplement to mammography, it improves imaging sensitivity at the cost of reduced specificity and increased biopsy rates Berg et al.; Wang (2008; 2017).

Figure 2.4: Breast ultrasound examination (source: https://www.radiologyinfo.org/en/info/breastus).

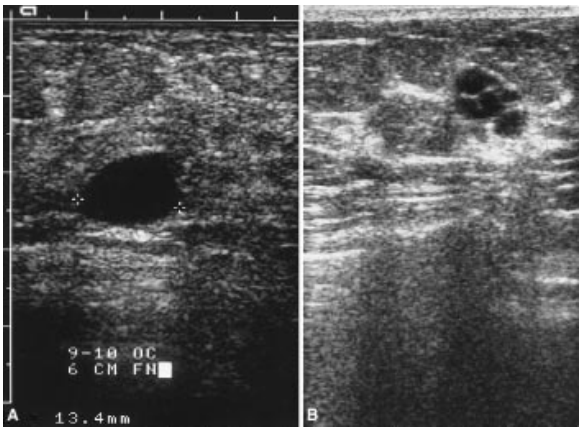

Figure 2.5: A. Simple cyst meeting all ultrasound criteria. B. Cluster of cysts (Kossoff, 2000).

Magnetic Resonance Imaging

Breast MRI is an imaging technique that uses magnets and radio waves to produce three-dimensional, detailed images of breast anatomy based on the magnetic property of the hydrogen nucleus that is abundant in the water that constitutes living tissue.

During breast MRI (as shown in Fig. 2.6), patients normally require an intravenous injection of contrast into the arm before or during the examination to help create clearer images and outline abnormalities more easily. The patient is usually lying face down on a padded platform with cushioned breast openings. Each opening is surrounded by a breast coil, which is a signal receiver that works in conjunction with the MRI unit to produce images. The platform then slides into the center of the tubular MRI machine. The technician looks at the MRI through a window while monitoring for any potential motion.

Figure 2.6: Breast magnetic resonance imaging (MRI, source: https://www.mayoclinic.org/diseases-conditions/breast-cancer/diagnosis-treatment/drc-20352475).

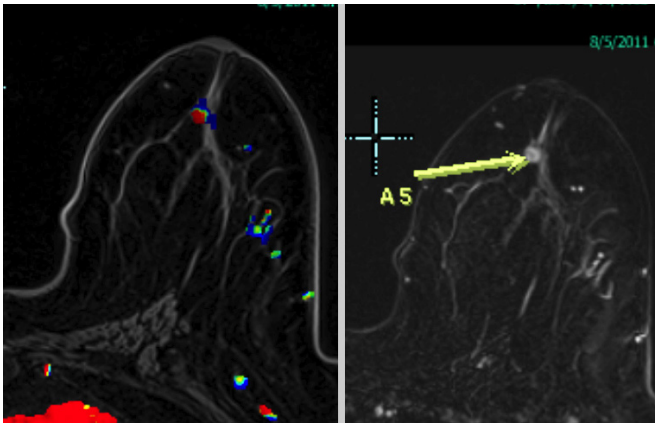

Figure 2.7: Bulky left axillary lymphadenopathy on MRI image. The arrow indicates a 10-mm enhancing retro areolar mass (Shahid et al., 2016).

Breast MRI can be used in all aspects of breast cancer treatment, including monitoring treatment response, monitoring high-risk patients, assessing breast cancer metastasis, and studying tumor recurrence (Mann et al., 2008). In particular, it has advantages over mammography and ultrasound in the detection of axillary lymphadenopathy (as

shown in Fig. 2.7). It has been found to detect more than 60% of occult breast malignancies that accounts for less than 1% of all breast cancers (De Bresser et al., 2010). It also reduces total mastectomy due to its high sensitivity and low false-negative rate in lesion detection (Shahid et al., 2016). The preoperative evaluation of patients with newly diagnosed malignancies is another indication. It can detect breast masses with a sensitivity approaching 100% and help determine chest wall and pectoral muscle involvement. Breast MRI is particularly beneficial in the preoperative evaluation of women with a pathologic diagnosis of invasive lobular carcinoma (ILC). It can greatly improve screening in certain high-risk groups (Schnall, 2000) and has higher spatial and temporal resolution and a better signal-to-noise ratio (Lehman and Schnall, 2005). However, it is not suitable for the general population due to its high false positive rate, expensive cost, time-consuming nature, insufficient units, requirements for professional experience, and insufficient clinical utility.

Positron Emission Tomography–Computed Tomography

Positron emission tomography–computed tomography (better known as PET-CT or PET/CT) is a nuclear medicine imaging technique that combines a positron emission tomography (PET) scanner and an x-ray computed tomography (CT) scanner in one gantry. It generates a sequence of images from both devices in the same session and combines them into a single superposed (co-registered) image. This combination aligns (or correlates) PET-obtained spatial distribution of metabolic or biochemical activity in the body with CT-obtained anatomic features.

Figure 2.8: PET/CT Scan for Breast Scanning (source: https://www.saintjohnscancer.org/breast/breast-health/breast-evaluation/other-tests).

During the PET/CT scan (as shown in Fig. 2.8), patients usually lie on an exam

table and are inserted with an intravenous (IV) catheter into the vein of one hand or arm when necessary. The injected radiotracer normally takes 30-60 minutes to travel through the body and be absorbed by the region under examination. After moving into the PET/CT scanner, patients are scanned by the CT first and then by the PET. A second CT scan with intravenous contrast is occasionally needed. A total scan takes around 30 minutes.

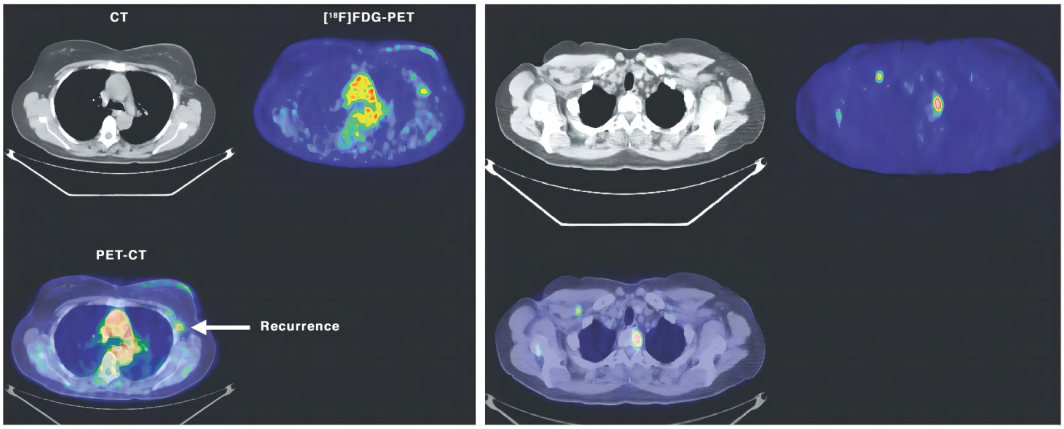

Figure 2.9: Breast cancer PET/CT images with recurrence in the left and metastases in the right (Zangheri et al., 2004).

PET/CT Antoch et al.; Pelosi et al. (2003; 2004) shows great performances in the classification of malignant and benign lesions, staging and re-staging disease, and making therapy plans, especially the one based on the widely used radiotracer glucose analogue 2-[18F]fluoro-2-deoxy-D-glucose (18F-FDG). It exhibits a high diagnostic accuracy in recurrent (Fig 2.9-left) or metastatic breast cancer (Fig 2.9-right) (Moon et al., 1998) and is particularly applicable when the only indicator of recurrence is a rise in serum tumor markers (Suarez et al., 2002). PET/CT can be used in re-staging and establishing the correct management of breast cancer patients both during early re-staging after primary treatment and during follow-up (Zangheri et al., 2004). It may provide important prognostic information regarding disease-free and overall survival in patients with locally advanced breast cancer (Avril et al., 1999). PET/CT is also useful in monitoring breast cancer response to chemotherapy and can improve the accuracy of treatment response evaluation (Zangheri et al., 2004). It shows a diagnostic sensitivity of around 90% and a specificity from 83% to 100% in detecting primary breast cancer and classifying malignant from benign disease (Avril et al., 1996; Scheidhauer et al., 1996; Buck et al., 2002). However, its expensive cost and limited spatial resolution discourage its application in the screening and diagnosis of primary breast tumors.

Microscopic Imaging: Biopsy

Breast biopsy is a breast cancer examination procedure in which a small sample of breast tissue is taken by surgical methods such as local excision, needle puncture aspiration, scraping, and removal and then sectioned for microscopic investigation.

During the biopsy procedure, an injection is first needed to numb the area of the breast to be biopsied. There are different ways to achieve breast biopsy, i.e., fine-needle

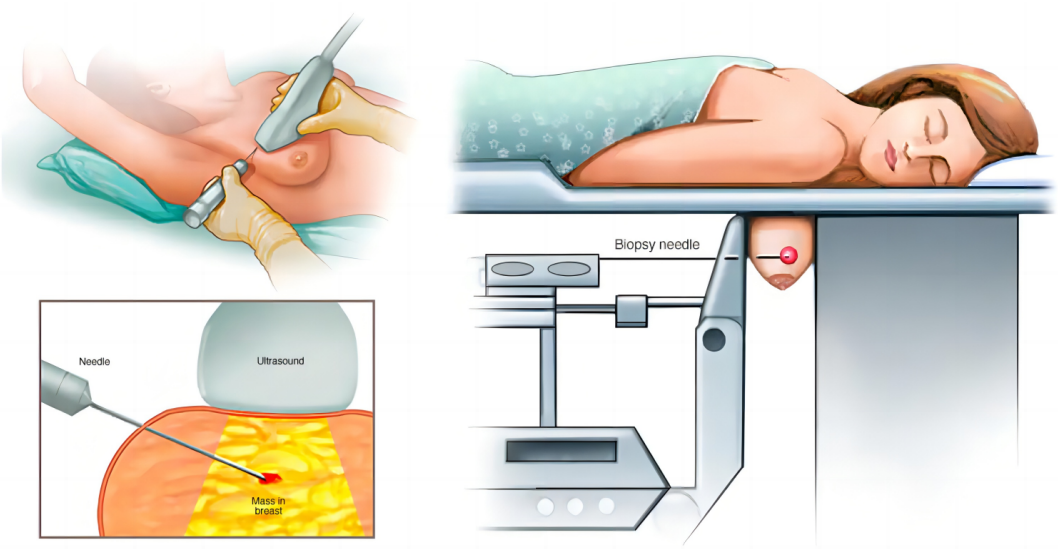

Figure 2.10: Breast biopsy (source: https://www.mayoclinic.org/tests-procedures/breast-biopsy/about/pac-20384812).

aspiration biopsy (Amedee and Dhurandhar, 2001), core needle biopsy (Heslin et al., 1997), stereotactic biopsy (Huang et al., 2014), and surgical biopsy (Kasraeian et al., 2010). Fine-needle aspiration biopsy is the simplest breast biopsy and can be used to evaluate a lump felt during a clinical breast exam. For the procedure, patients lie on a table with a doctor using one hand to steady the lump and another to insert a very thin needle into it. The needle is connected to a syringe that can collect a sample of cells or fluid from the lump. This method is able to quickly achieve the distinction between a fluid-filled cyst and a solid mass. It may also help avoid a more invasive biopsy procedure. Core needle biopsy is a biopsy procedure in which a radiologist or surgeon uses a thin and hollow needle to take tissue samples from the breast mass under the guidance of ultrasound (Apesteguía and Pina, 2011) or MRI (Lilly et al., 2020) (depending on the location of the mass). This biopsy collects and analyzes several samples approaching the size of a grain of rice. Take the ultrasound-guided core needle biopsy as an example (as shown in Fig. 2.10-left). For this procedure, patients lie on their backs or sides on an ultrasound table. A doctor or radiologist uses one hand to hold the ultrasound transducer against the breast to locate the mass and another to insert the needle inside to take several core samples of tissue. Stereotactic biopsy uses mammograms to pinpoint suspicious areas within the breast. For this procedure (as shown in Fig. 2.10-right), patients generally lie face-down on a padded biopsy table with one of their breasts positioned in a hole and firmly compressed between two plates. Then a radiologist makes a small incision, about 0.25 inches long, into the breast, followed by inserting a needle or vacuum-powered probe to obtain samples. Surgical biopsy is usually done in an operating room with some or all breast mass removed for examination, where a wire or seed localization technique may be necessary to map the mass route. For wire localization, a thin wire tip is positioned within or through the breast mass. For seed localization, a small radioactive seed is placed inside the breast to guide the surgeon. For this procedure, the removed tissue is sent to the hospital lab for evaluation to confirm whether breast cancer is present in

the mass and its margins (positive margins). Fig. 2.11 shows an example of ILC in the histopathological image obtained during the biopsy.

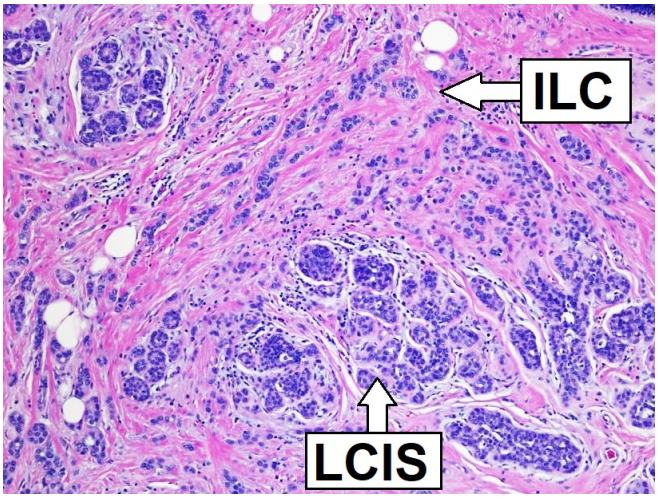

Figure 2.11: histopathological image of invasive (or infiltrating) lobular carcinoma (ILC) obtained during the biopsy (Wikimedia, 2020).

Breast biopsy (O'Flynn et al., 2010) is adopted to solve the diagnostic problem of impalpable breast lesions found during breast cancer screening. It is routinely performed to confirm suspicious of malignancy (category 4 findings) and highly suspicious malignancy (category 5 findings) lesions (Liberman and Menell, 2002). Apart from making decisions for the cancer situation (e.g., malignancy), a needle biopsy is also adopted to characterize the lesion histologically and to obtain information for making overall oncological treatment plans, which involves histological grade and type, basal subtype, hormone, and HER2 receptor status, and genetic profiling. However, breast biopsy faces risks like bleeding, formation of hematoma, and post-procedure pain.

2.1.3 Datasets for Investigating the CAD System of Breast Cancer Diagnosis

This thesis mainly focused on the investigation of the CAD System algorithm for breast cancer diagnosis with biopsy, which ultimately goes to the classification of histopathological breast cancer images. In order to evaluate and verify the effectiveness of our algorithm (i.e., deep learning models) in a variety of clinically relevant digital breast cancer histopathology tasks, two public datasets were used in our experiments. The first one is Camelyon16 patch-based dataset (Rony et al., 2019) that is derived from the Camelyon16 dataset (Bejnordi et al., 2017), and the second is the BreakHis dataset (Spanhol et al., 2016b).

Camelyon16 Patch-Based Dataset

The Camelyon16 patch-based dataset with images of 512×512 pixels is produced with the protocol designed by Rony et al. (2019) in previous work. Its source images are from the public Camelyon16 dataset (Bejnordi et al., 2017) that is used to detect metastases

in H&E stained tissue sections of sentinel auxiliary lymph nodes (SNLs) of women with breast cancer. It includes 270 whole-slide images (160 normal and 110 metastatic) with an average size of around 65000×45000 pixels for training and 129 images (80 normal and 49 metastatic) for testing. Specifically, the images in Camelyon16 patch-based dataset are sampled patches from the whole-slide images. As a result, it contains 24348 samples for training, 8858 samples for validation, and 15664 samples for testing. The ratio of normal images to metastatic images in all three sets satisfies 1:1. Each sample has both image-level and pixel-level annotations, and the latter has three cases, i.e., masks without any metastatic region, masks containing both normal and metastatic regions, and masks having only metastatic regions. Fig. 2.12 gives some examples of this dataset, and the red regions of the first row are metastatic.

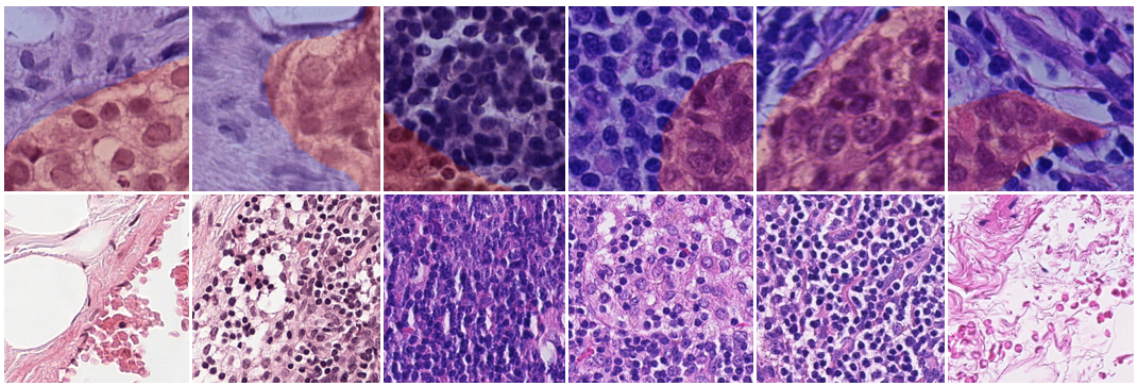

Figure 2.12: Examples of metastatic (upper row, red regions are metastatic) and normal (bottom row) images from Camelyon16 patch-based dataset.

BreakHis Dataset

The BreaKHis dataset has 7909 pathological images from 82 breast cancer patients, where 5429 images are malignant (from 58 patients), and the other 2480 images are benign (from 24 patients). Each image is of 3-channel RGB with a size of 700×460 pixels and eight-bit color depth in each channel. All the patients in the dataset have images acquired with objective magnifications of 40×, 100×, 200×, and 400× respectively. Additionally, both the benign and malignant images contain four sub-classes. The benign cases are adenosis (A), fibroadenoma (F), tubular adenoma (TA), and phyllodes tumor (PT). The malignant cases are ductal carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC), and papillary carcinoma (PC). Fig. 2.13 shows a typical breast malignant image with four magnification factors. The highlighted rectangular areas are the regions of interest (ROIs) to be magnified. Some examples of four sub-classes of benign and malignant images with a magnification of 40× are shown in Fig. 2.14. In our work, we restricted the task to binary classification, i.e., benign versus malignant, and used this dataset only for classification evaluation due to its lack of pixel-level annotations.

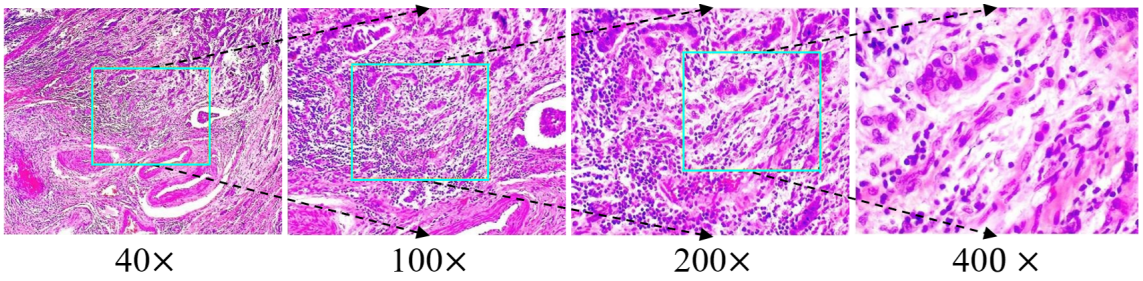

Figure 2.13: Slides of breast malignant tumor seen with different magnification factors.

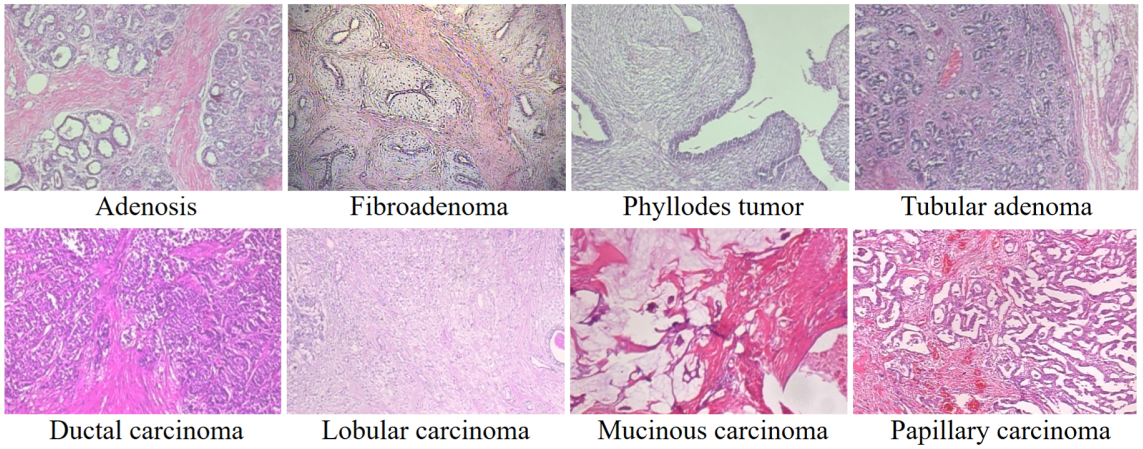

Figure 2.14: Examples of four benign sub-classes (upper row) and four malignant sub-classes (bottom row) with a magnification factor of 40×.

2.2 Display Panel Defect Detection

2.2.1 Display Panel Manufacturing

Array Process

The Array process of the LCD panel mainly contains film, yellow light, etch, and peel film. Since electrons are needed to drive the movement and alignment of the LCD molecules, conductive parts that control them are necessary on the TFT glass (the carrier of the LCD). Indium Tin Oxide (ITO) is one of the materials that can be used to achieve this purpose. ITO is transparent and thus can act as a thin-film conductive crystal that does not block the backlight. Similar to printing the circuit on the PCB board, ITO film requires drawing the conductive circuit on the entire LCD board. Fig. 2.15 provides the detail of the Array process, which is described as follows:

- Stage 1: The ITO film layer is deposited smoothly and uniformly on the TFT glass, which is cleaned by ionized water.
- Stage 2: Form a uniform photoresist layer on the ITO film-deposited glass. Bake it for a certain time to partially volatilize the solvent of the photoresist to make the photoresist material more adhesive to the ITO glass.

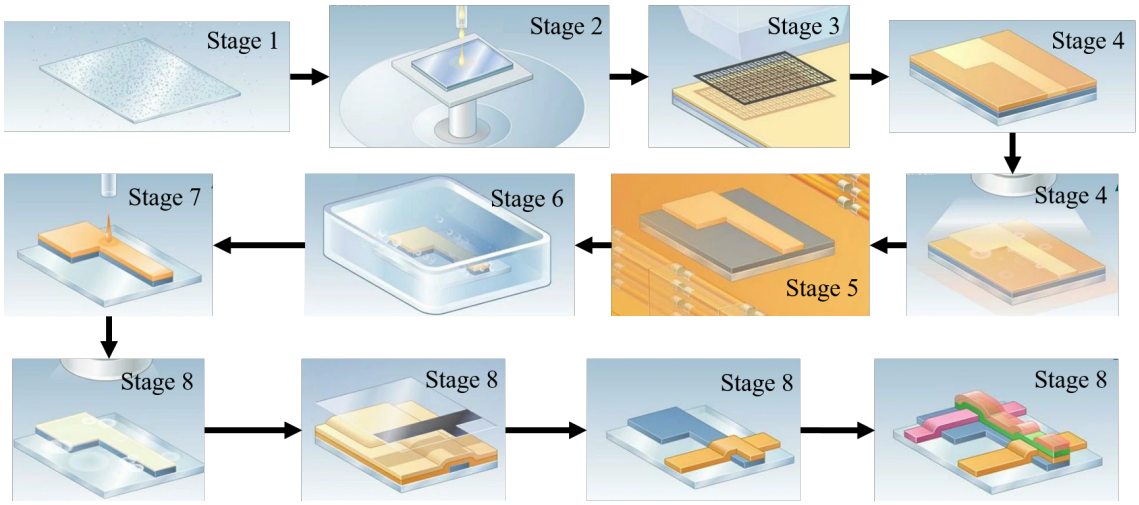

Figure 2.15: Array process

- Stage 3: Use ultraviolet light (UV) to illuminate the surface of the photoresist through a pre-made electrode pattern mask. It causes the photoresist layer to react and be selectively exposed to ultraviolet light by covering the photoresist on the glass coated with the photoresist.

- Stage 4: The light part is unexposed, and the dark part is exposed (taking a pixel unit as an example). Use the developer to wash away the exposed part of the photoresist leaving only the unexposed part, and the dissolved photoresist is then washed away with deionized water.

- Stage 5: Heat and bake the TFT glass after stage 4 to make the unexposed photoresist adhere more firmly to the ITO glass.

- Stage 6: Use the appropriate acid etching solution to etch off the ITO film part without the cover of the photoresist, leaving the ITO film part under the photoresist. Because ITO film is made of In2O3 and SnO2, it is easy to react with acid, and the photoresist can protect it.

- Stage 7: Use an alkali solution (NaOH solution) with high concentration (as a stripping solution) to peel off the remaining photoresist on the glass, which forms ITO graphics exactly consistent with the photolithography mask.

- Stage 8: Use an organic solution to rinse the basic label of the glass and remove the photolithographic tape after the reaction to keep the glass clean.

Cell Process

The LCD panel is structured like a sandwich, with the TFT glass underneath and the color filter on top. Thus, the terminal Cell process includes gluing the TFT glass to the top and the colored filter to the bottom. Fig. 2.16 shows the detail of the Cell process, which is introduced as follows:

- Stage 1: Use ionized water to rinse the TFT glass after the Array process.
- Stage 2: Uniformly coat the organic polymer directional material on the surface of the glass.
- Stage 3: Rub the surface of the layer in a specific direction with the flannelette material. It makes the LCD molecules arranged along the friction direction of the aligned layer in the future, thus ensuring the consistency of the alignment of LCD molecules. Clean the TFT glass substrate and wash away contaminants such as flannelette thread.
- Stage 4: Apply sealant coating to conglutinate the TFT glass substrate to the color filter and to prevent LCD outflow.
- Stage 5: Coat color filters with an orientation film.
- Stage 6: Align the alignment film fixed on the surface of the filter.
- Stage 7: Spray a pad on the color filter surface to keep a distance between it and the TFT glass substrate.
- Stage 8: Enter the TFT glass substrate process again and inject the LCD into the sealant frame coated on the TFT glass substrate.
- Stage 9: Apply the conductive adhesive to the frame in the glass bonding direction of the color filter to ensure external electrons flow into the LCD layer. Bond two pieces of glass together according to the bonding mark on the TFT glass substrate and the color filter. Solidify the bonding material at high temperatures to make the upper and bottom glasses fit statically.
- Stage 10: Cut the LCD plate according to the designed size. Place a polarizer on both sides of each LCD substrate, with the horizontal polarizer facing outwards and the vertical one facing inwards.

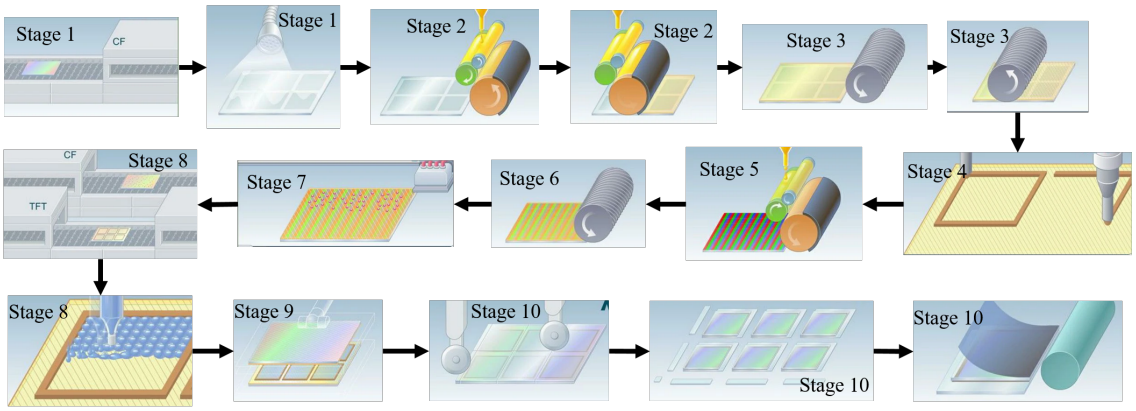

Figure 2.16: Cell process

Module Assembly Process

The Module Assembly process is mainly the press fit of the drive IC and the LCD substrate and the integration of the printed circuit board. This process can transmit the display signal received from the main control circuit to the drive IC to drive the LCD molecules to rotate and display the image. In addition, the backlight part will be integrated with the LCD substrate at this stage, and the LCD panel is finally completed. Fig. 2.17 shows the detail of the Module Assembly process, which is described as follows:

- Stage 1: Press the hetero-conductive adhesive on the two edges produced in the Cell process, which acts as a bridge that allows external electrons to enter the LCD substrate layer.
- Stage 2: Press a drive IC on the LCD substrate, which outputs the required voltage to each pixel and controls the torsion degree of the LCD molecules.
- Stage 3: Use the hetero-conductive adhesive to glue together the end of the flexible circuit board and the printed circuit board.

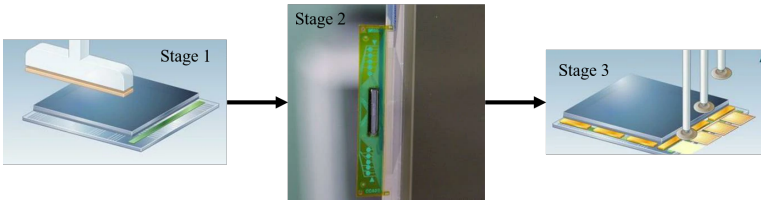

Figure 2.17: Module Assembly

2.2.2 Online Defect Inspection Algorithms

LCD and OLED panels have periodic textured surfaces at the Array stage. Traditional methods such as edge detection and threshold segmentation are not effective in detecting defects in such backgrounds where the pixel value of the defect is close to that of the background texture. The current optical inspection methods (i.e., AOI algorithms) for display panel manufacturing are mainly three-fold: differential, transform, and statistical.

Differential Method

The differential method is a simple and straightforward detection method. Its core idea is to obtain the residual image by performing the differential operation between the defect-free template image and the image to be detected and then determine the defects according to segmentation algorithms. The method mainly includes the generation of defect-free template images, the alignment of templates and defect images, and the extraction and determination of defects. Many works have been done to improve the performance of this method.

A hybrid inspection system incorporating differential and optical Fourier filtering was proposed (Nakashima, 1994), which is capable of detecting 8 types of defects. A line-scan camera was proposed (Kim et al., 2001) to replace the conventional surface array camera, where a quad-core parallel DSP was used to process the acquired images. This method improved the inspection speed and the inspection resolution. A polynomial fitting approach based on the differential was proposed (Baek et al., 2004) to reconstruct the background template image. This method eliminated the need to generate and store templates before detection and reduced the difference between the template and the defect image caused by illumination factors and environmental noise. An image alignment method based on local template matching was proposed (Su et al., 2008) to solve the problem of TFT-LCD defects at the micron level, which enables image alignment at the sub-pixel level and rapid automatic defect detection at the micron level.

The differential algorithm is simple, straightforward, and can achieve fast and effective detection of small defects, and thus it is widely used in surface defect detection. However, in order to eliminate the influence of background texture, the method often requires the introduction of template images, which are computationally complex to generate and time-consuming to align. It requires a large amount of storage capacity for the template images and demands template alignment accuracy as well as image acquisition quality. The template-free differential method has a relatively greater potential for development.

Transform Method

The transform method first transforms the image signal from the spatial domain to the spectral domain utilizing Fourier transform, wavelet transform, Gabor transform, or discrete cosine transform. After that, the frequency components of the repetitive texture background are filtered out in the spectral domain, and the local anomalous information of the defects is retained. Then, the image reconstruction is performed by the corresponding inverse transform to obtain an image without repetitive texture. Finally, image defect determination is performed by image segmentation.

Global 1D Fourier transform and 2D Fourier transform methods were proposed (Tsai and Hung, 2005; Tsai et al., 2007) to achieve the elimination of periodic texture backgrounds in images at low and high resolutions. The parameters in the algorithms depend on experiment decisions, and the non-uniformity of gray-scale values has a large impact on background filtering. The independent component analysis method was proposed (Tsai and Lai, 2008) for defect detection. This method used the solution mixture matrix extracted from the defect-free 1D image to recover the detected 1D image and normalized the correlation to determine the location of the defect by calculating the similarity between the two. Due to the complex mathematical transformations applied in this method, it cannot meet the real-time detection requirements. Gabor filter and Principal component analysis (PCA) methods were proposed (Bissi et al., 2013) for the detection of defects under complex textures. This method overcomes the drawback that the Fourier transform cannot be localized for analysis and can extract relevant features in the direction of different scales in the spectral domain. Its time-frequency window size and shape are not adaptive, and the non-orthogonality

brings redundancy between different feature components. The 2D DFT-based defect detection algorithm was improved (Zhang et al., 2016) for defect detection. It utilized the Hough transform to detect high-energy spectral-domain lines of textures, which can have a good filtering effect on directional linear textures.

The transform method is a common means of periodic background filtering and fast detection, and it has rotation and scale invariance. It is suitable for line detection and has a good detection performance for defects of large sizes. Nevertheless, the transform method can usually only determine the presence or absence of defects while not providing more detailed information for the classification of defects. The performance of spectral-based reconstruction greatly influences micro defect detection accuracy, and tiny defects are not easily detected.

Statistical Method

The statistical method performs feature extraction or data dimensionality reduction based on the acquired image data. The extracted feature information, such as texture feature, geometric feature, and gray-scale feature, is then input into the classifier to complete the determination of the presence or absence of defects or the recognition of defect classes. There are many related research works in both feature extraction and classifier construction.

A saliency model-based defect detection approach was proposed (Lee et al., 2004) to detect fuzzy defects. The method built saliency models by using color (gray-scale images can be replaced by periodicity), intensity, and orientation information of the input image and performed anisotropic filtering using the LoG operator. Fuzzy theory was introduced (Zhang and Zhang, 2005) into the pattern recognition system to construct a fuzzy expert system. The fuzzy theory was combined with neural networks to construct an FNN classifier. It enables the classification and grading of the defect. However, the defect classification of this method relies too much on the formulation of fuzzy rules and lacks specific theoretical guidance. Principal component analysis (PCA) and linear discriminant analysis were proposed (Kang et al., 2009) to perform feature dimensionality reduction on the image input to the classifier to improve the classification speed of the classifier while reducing the sensitivity to noise. This method can only perform dimensionality reduction for linear data of images. A fuzzy support vector data description algorithm based on kernel fuzzy c-means was proposed (Liu et al., 2009) to build single-class classification. This method used four features of texture entropy, energy, contrast, and uniformity for classifier training. Its defect detection accuracy can reach 99% and classification accuracy 96%, with a fast detection speed suitable for online real-time defect detection and classification. However, this method can only detect defects without the circuit texture part inside each cycle. Methods based on the combination of brightness distribution, linearity, and morphological characteristics of the defect image were proposed (Noh et al., 2009) to separate 5 types of commonly occurring defects. Linear fitting, multi-level brightness, and shape morphology were also used in these methods. A method based on faster regions with convolutional neural networks (Faster R-CNN) was proposed (Lei et al., 2018a) to identify and locate defects of polymeric polarizer in TFT-LCD panels.

Statistical methods based on deep learning can extract and learn the features of

image data more comprehensively, which can therefore have more accurate detection performance in defect classification compared with traditional differential and transform methods. However, because the deep learning-based algorithm needs a large amount of data training and has high computational complexity, it also leads to an advanced requirement for the hardware of the detection equipment that runs the algorithm.

2.2.3 Dataset for Investigating the AOI System of Online Defect Detection of Display Panels

We used a display panel defect dataset collected through a microscope system in real-world industrial manufacturing to evaluate and verify the effectiveness of our method. The dataset contains 571 defective images of 1024×768 pixels, and the images present six classes of backgrounds. Each of them has different types of defects, such as the foreign object, film off, film adhesive, malformation, etc., as shown in Fig. 2.18.

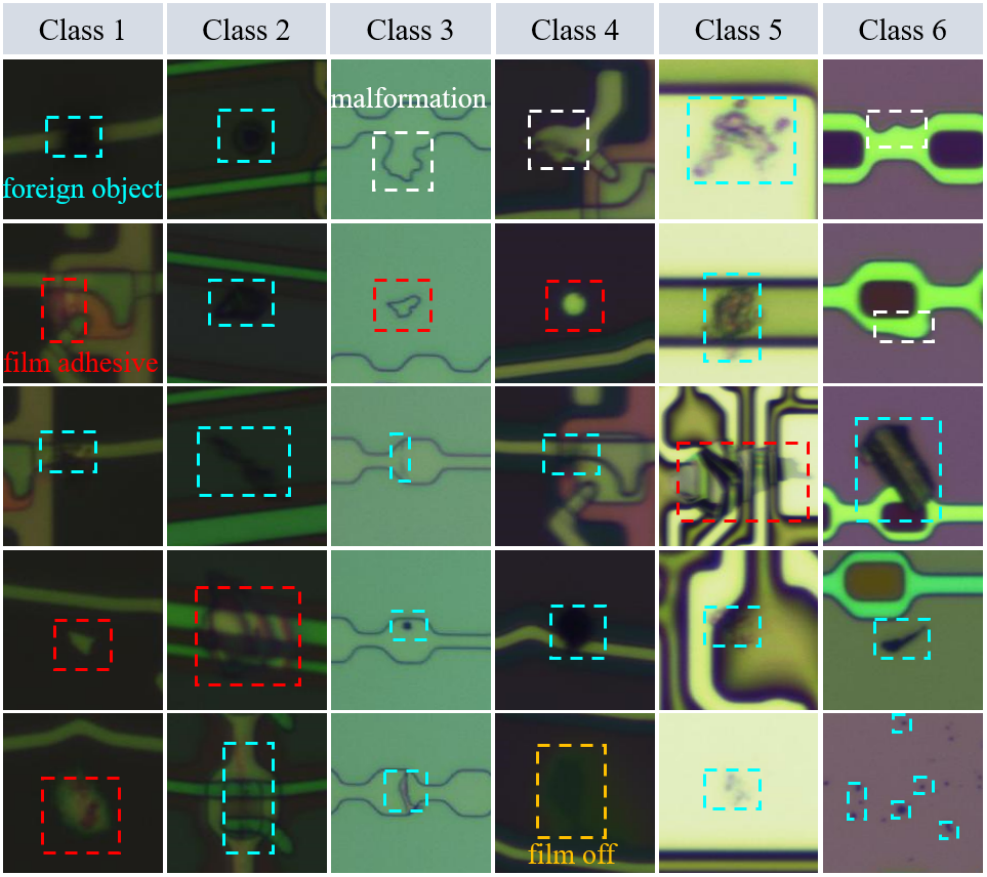

Figure 2.18: Different defects in six classes of backgrounds of display panels.

The publicly available tool LabelMe was used to conduct the pixel-wise annotation to obtain the ground truth mask of our task. A dataset with a large number and variety is of utmost importance for the outstanding performance of the deep learning method in defect detection. To meet this requirement and surmount over-fitting, we performed data augmentation on the defective images. The images were enhanced through

Table 2.1: Distribution of training and test sets after data augmentation.

Classes	1	2	3	4	5	6	Total
Training	2109	2811	2304	2400	2304	2208	14136
Testing	91	93	144	96	112	108	644

random cropping, rotation (90^o, 180^o, and 270^o), flip, and brightness adjustment. The final distribution of the images in the training and testing sets is given in Table 2.1.

2.3 State-of-the-Art Deep Learning Methods

This thesis is dedicated to developing a configurable convolutional neural network to achieve both the explainable classification of breast cancer at the decision-making stage of clinical biopsy and online defect detection of display panels. Its key techniques lie in convolutional neural networks and weakly-supervised learning, with the involvement of classification and segmentation (or localization) needs. Therefore, a comprehensive introduction to the related technical background is necessary.

2.3.1 Convolutional Neural Networks (CNNs)

Convolutional neural networks (CNNs) are feedforward neural networks with a deep structure, which perform mathematical convolution operations on the input image through convolutional kernels (weight matrix) of certain sizes and numbers.

Hubel and Wiesel's research (Hubel and Wiesel, 1962, 1968) during 1950-1960 found that the visual cortices of cats and monkeys contain neurons responding to a small visual area, respectively. When their eyes were held still, a visual stimulus in a certain area excited a single neuron, and that area is called the receptive field of that neuron. Its neighboring cells have similar and overlapping receptive fields. To form a complete visual image, the size and location of the neurons' receptive field across the visual cortex vary systematically. In 1980, Fukushima proposed (Fukushima and Miyake, 1982) the concept of neocognitron based on the receptive field, which can be regarded as the predecessor of CNN. The neocognitron decomposes a visual pattern into many sub-patterns (i.e., features) and then processes them in a hierarchically connected feature plane. It tries to model the visual system to recognize objects even when they are displaced or slightly deformed.

Technically, the first CNN model was proposed by LeCun et al. in 1989 (LeCun et al., 1989) and improved in 1998 (LeCun et al., 1998), which is called LeNet-5 aiming to recognize handwritten digits. Fig. 2.19 shows the architecture of this CNN model, which contains one input layer, two convolutional layers, two pooling layers, and three fully connected layers (the last one is the output layer). with the training based on the backpropagation algorithm (Rumelhart et al., 1986; Hecht-Nielsen, 1992), it can effectively extract target representations of the raw input image, which makes it capable of recognizing visual patterns from pixels without preprocessing.

Due to the good performance of LeNet-5 in handwritten digit recognition, it received the attention of Alex Krizhevsky, whose AlexNet (Krizhevsky et al., 2012) with

a deeper CNN proposed in the "ImageNet Large Scale Visual Recognition Challenge" event (a.k.a. "ImageNet competition") in 2012 was able to classify 1.2 million high-resolution images from 1000 categories and won the championship. Fig. 2.20 provides the architecture of AlexNet model, which contains one input layer, 5 convolutional layers, three pooling layers, and two fully connected layers. Compared to LeNet-5, it has three more (deeper) convolutional layers, using ReLU activation function to replace Sigmoid function (the former is more efficient in computation and can mitigate vanishing gradient problem), using dropout strategy to control the model complexity of the fully connected layer and reducing over-fitting, and using local response normalization (LRN). The LRN creates a competition mechanism for the activity of local neurons so that the values with larger responses among them become relatively larger and neurons with smaller feedback are inhibited, which enhances the model generalization. Since the AlexNet was proposed, CNNs have received widespread attention and spurt in development.

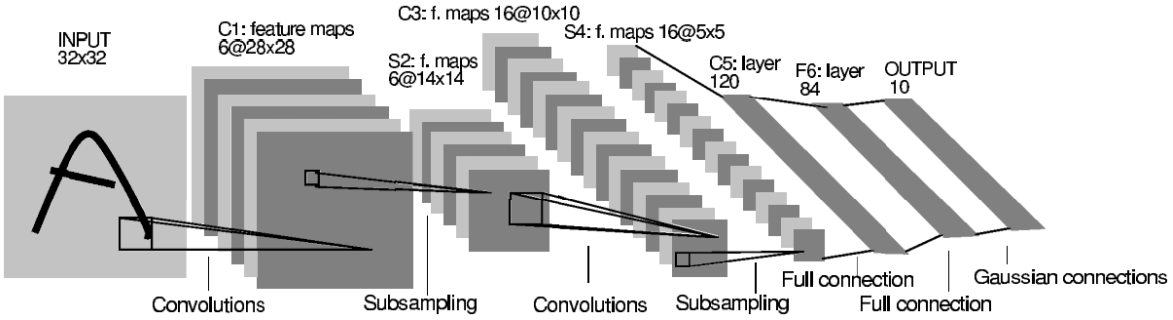

Figure 2.19: Architecture of LeNet-5 (LeCun et al., 1998). Each plane is a feature map.

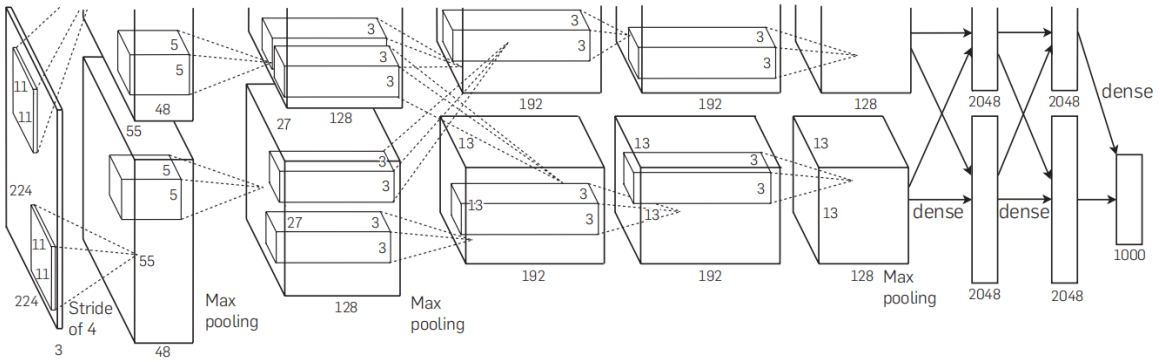

Figure 2.20: An illustration of the architecture of AlexNet (Krizhevsky et al., 2012), explicitly showing the delineation of responsibilities between the two GPUs.

In 2014, Simonyan and Zisserman proposed the VGG model (Simonyan and Zisserman, 2014a) with a depth of 11 to 19 layers. Fig. 2.21 shows the architecture of VGG16 having 16 layers. It replaced the large-size convolution kernel (11×11) of AlexNet with many small kernels of size 3×3, which has the same receptive field as the large one and increases the number of weight layers and nonlinear units, thus enhancing the network performance.

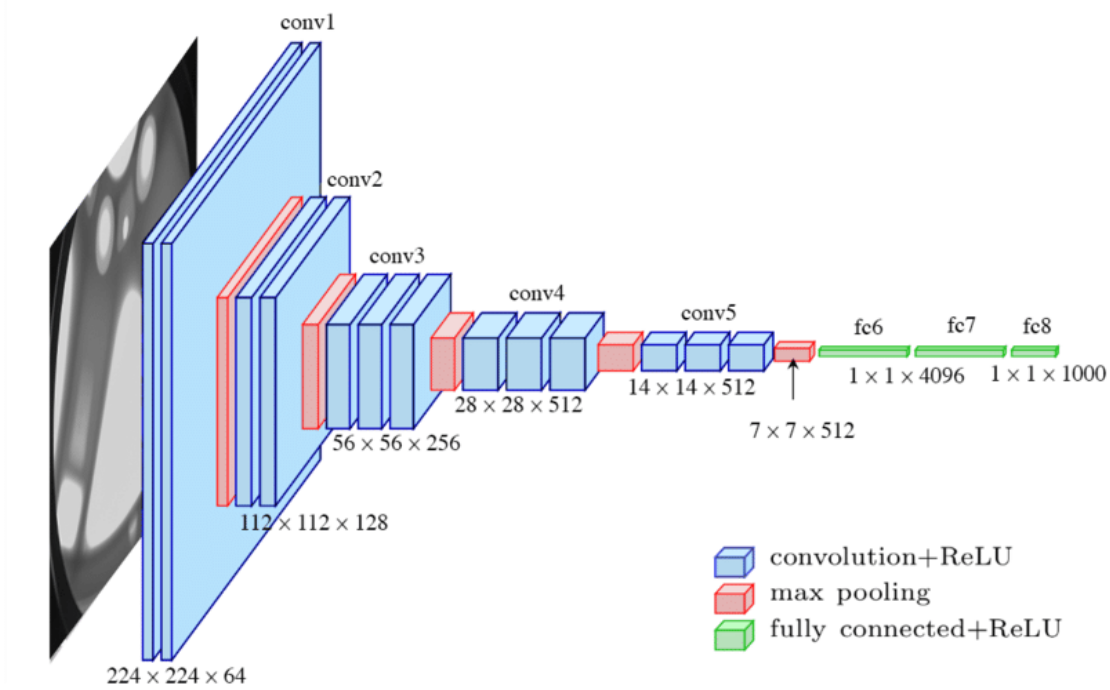

Figure 2.21: Architecture of VGG16 (Ferguson et al., 2017).

Later in 2015, Szegedy et al. developed a GoogLeNet (Szegedy et al., 2015a) that has a much deeper architecture, and they further improved it into Inception-v3 (Szegedy et al., 2016) in 2016 and Inception-v4 and Inception-ResNet (Szegedy et al., 2017) in 2017, respectively. Fig. 2.22 gives the architecture of the first version, i.e., GoogLeNet. They developed a seminal basic neuronal structure called Inception, which consists of convolutional kernels of different sizes (1×1, 3×3, and 5×5) concatenated in parallel and fuses features at different scales. The 1×1 convolution kernel is another bright spot that reduces feature channels with a single weight for each of them and thus reduces computational complexity. GoogLeNet is a dense sparse structure formed by cascading several such Inception structures, which enables the clustering of sparse matrices into dense submatrices so as to improve computational performance.

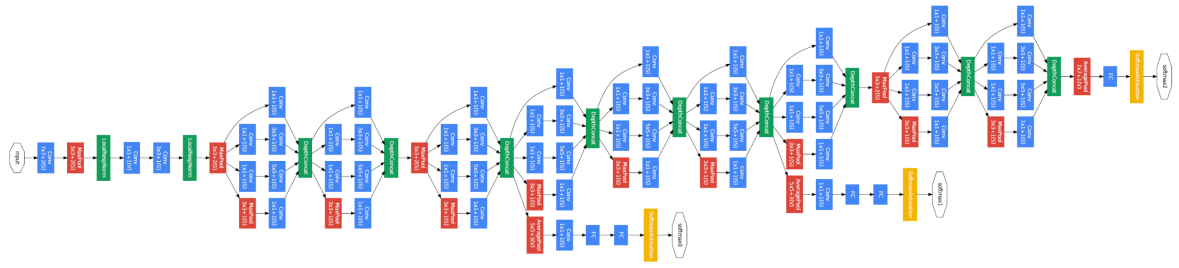

Figure 2.22: Architecture of Inception-v1 (a.k.a. GoogLeNet) (Szegedy et al., 2015a).

With the increase of the network depth, its training becomes more complex due to the internal covariate shift phenomenon (i.e., the input distribution of each layer

changes with the parameters of its previous layer). This phenomenon makes it necessary to set a lower learning rate and put more requirements for the parameter initialization, which unavoidably burdens the training, especially for networks with saturating nonlinearities. Therefore, Ioffe and Szegedy proposed (Ioffe and Szegedy, 2015) batch normalization one month after the publication of GoogLeNet. It solves the covariate shift phenomenon and allows the network to be trained with a larger learning rate and a coarse parameter initialization. Since then, researchers have used batch normalization as an essential element of CNNs, usually immediately after the convolutional layer and before the activation function.

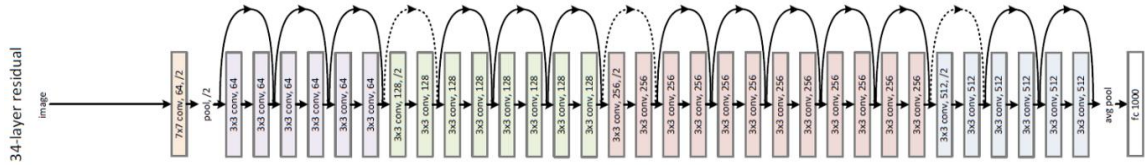

Figure 2.23: Architecture of ResNet-34 (He et al., 2016a).

In 2016, He et al. developed deep residual learning and ResNet (He et al., 2016a) with 18 to 152 layers to solve the degradation problem, i.e., the accuracy of a model often decreases once its depth is increased to a certain degree by the simple stacking of convolutional layers. Fig. 2.23 provides the architecture of the ResNet-34 with 34 layers. It added a shortcut connection between the input and output of each residual block to achieve feature identity mapping. It also used the 1×1 convolution layer to reduce computational complexity, where residual learning is extremely necessary. With the addition of such residual structure, the degradation problem in a deep CNN is considerably solved.

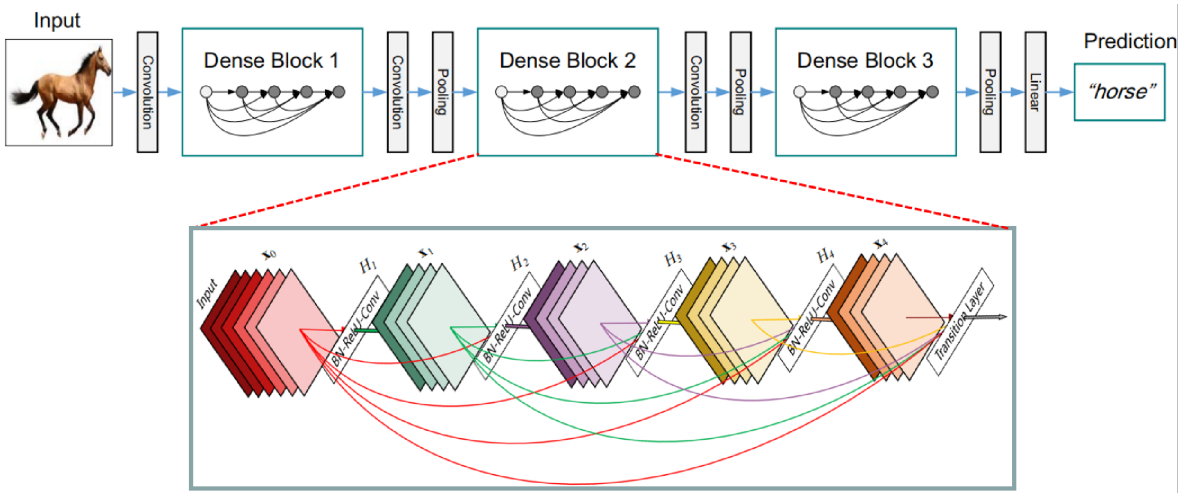

Figure 2.24: Architecture of DenseNet (Huang et al., 2017).

One year later, Huang et al. proposed the DenseNet (Huang et al., 2017) with much more information flowing between layers (compared with ResNet) and a dense connectivity pattern to solve the vanishing gradient problem. They took the connectivity pattern between different layers further to the highest level that concatenates all the

output features of the previous layers along with the current layer output features and feeds them to the subsequent layer for each layer, as shown in Fig. 2.24, which enables every layer to obtain the "collective knowledge" coming from the previous layers. This feature reuse development takes full advantage of the features learned at each layer, which avoids the re-learning of redundant feature maps and reduces the requirement for more parameters, thus increasing computational efficiency. In addition, feature reuse helps reduce over-fitting and makes it easier for the network to converge during the training compared to previous CNNs due to its implicit deep supervision nature that each layer of DenseNet has straight access to the gradients from the loss function and the network input.

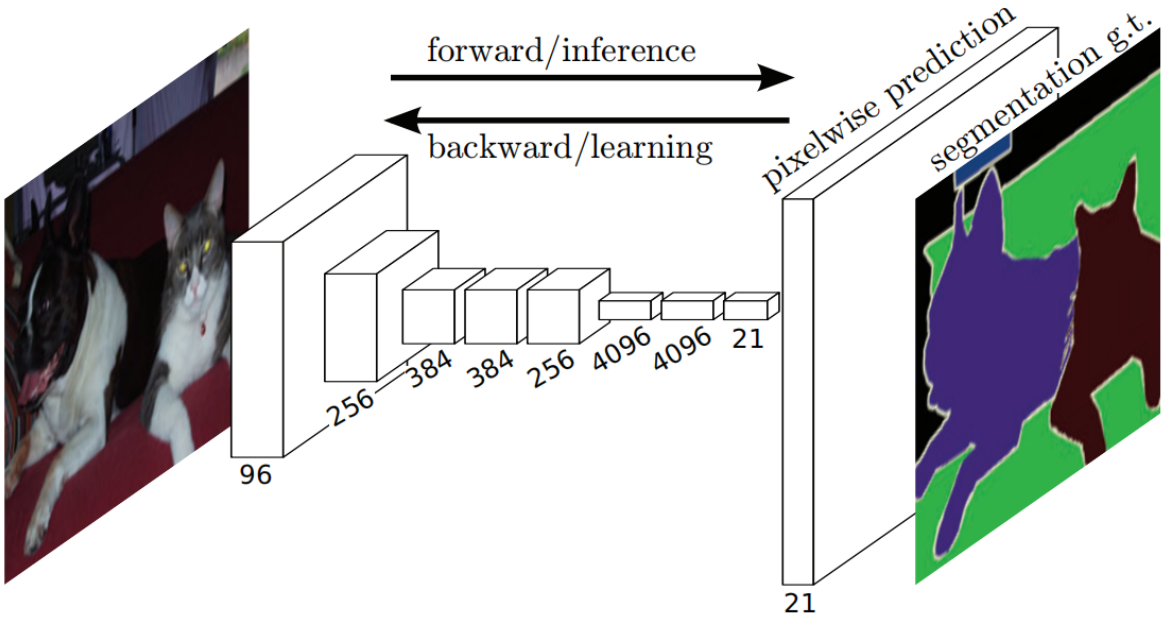

Figure 2.25: Architecture of FCN (Long et al., 2015).

Noteworthy, all the above CNNs are developed to solve classification problems, i.e., they only output the probabilities corresponding to the categories. However, in many cases, it is required to obtain the position (such as target recognition) or the whole contour of the target object (such as semantic segmentation), and this is the localization or segmentation problem that is a current research hotspot. The first CNN model for segmentation tasks that received widespread attention and is regarded as the pioneer is the FCN (i.e., fully convolutional network), which was proposed (Long et al., 2015) by Long et al. in 2015. Fig. 2.25 depicts the architecture of the FCN model. It replaced all the fully connected layers of the previous classification CNNs with convolutional layers of the same dimensional sizes, which enables the CNN to learn features of input images with large sizes and extract information relevant to the target localization and boundary. It then added up-sampling layers to recover the resolution of the final output to that of the network input through deconvolution, i.e., the reverse of convolution. The recovered map is the result indicating the localization or segmentation. Furthermore, the FCN model utilized a method called skip to combine knowledge of deep and shallow layers to make a dense prediction, which greatly improved the segmentation

accuracy.

Just five months later, Ronneberger et al. improved the FCN model to a CNN called U-Net (Ronneberger et al., 2015) that is more applicable to medical image segmentation with a training dataset of small size. Fig. 2.26 gives the architecture of the U-Net. U-Net used approximately the symmetric structure of the down-sampling convolutional layers (except the down-sampling was replaced with up-sampling) as the up-sampling path to pass on context information to resolution recovering layers. The skip strategy was extensively improved to reuse all the knowledge from the down-sampling path by concatenating features output from each layer (except the last layer) of the up-sampling path with that of the symmetric layer in the down-sampling path. In this way, U-Net achieved an average IoU of 77.5% in the segmentation of HeLa cells.

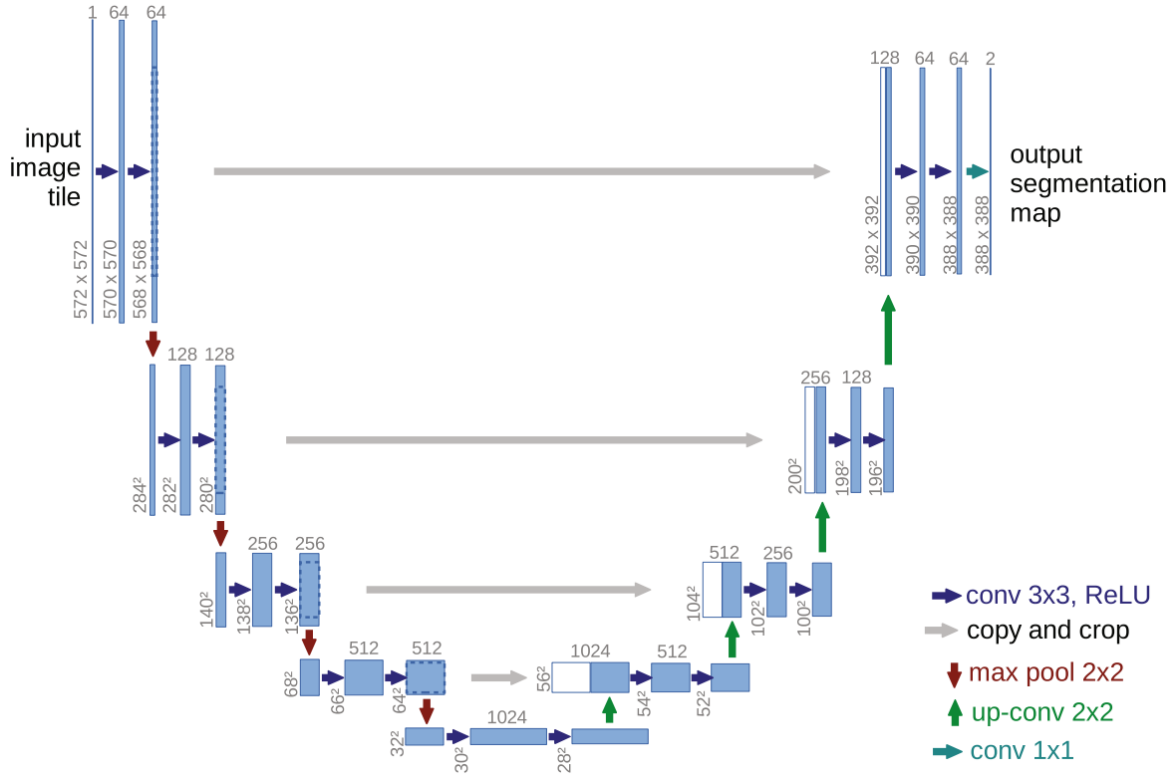

Figure 2.26: Architecture of U-Net (Ronneberger et al., 2015).

In 2017, Badrinarayanan proposed a typical encoder-decoder segmentation architecture termed SegNet (Badrinarayanan et al., 2017) for semantic pixel-wise segmentation of natural objects. Fig. 2.27 provides the archtecture of the SegNet. SegNet used the feature extraction module (i.e., all the convolutional layers) of VGG16 as the encoder (i.e., down-sampling path), and the decoder (i.e., up-sampling path) has the structure topologically identical to the encoder. Its key novelty is that the max-pooling indices calculated during the encoder were saved to up-sample feature maps of the symmetric decoder layers, which makes each pixel of the low-resolution feature map tend to revert to its original position in the high-resolution feature map and thus enhances the segmentation performance. Due to the great success of U-Net in medical image segmentation, researchers have studied it in more depth and have come up with many

improved versions, such as R2U-Net (Alom et al., 2018), Attention U-Net (Schlemper et al., 2019), MultiResUNet (Ibtehaz and Rahman, 2020), Multi-Res-Attention U-Net (Thomas et al., 2020), R2AU-Net (Zuo et al., 2021), Sharp U-Net (Zunair and Hamza, 2021), and Ds-transunet (Lin et al., 2022).

Among all the U-Net series, the Attention U-Net (Schlemper et al., 2019) proposed by Schlemper et al. in 2019 is the most classic CNN model introducing the attention mechanism into the segmentation network for medical images. Fig 2.28 shows the architecture of the Attention U-Net and its attention module. They developed an additive attention gate with the deep features from the decoder as the gate vector (has fewer

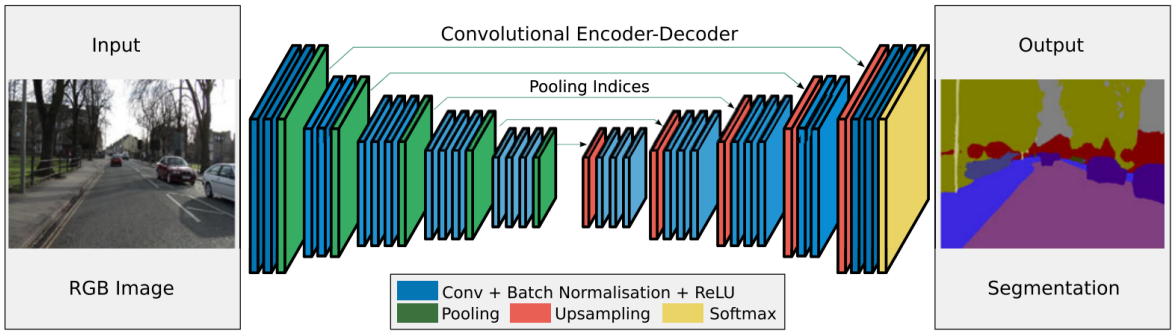

Figure 2.27: Architecture of SegNet (Badrinarayanan et al., 2017).

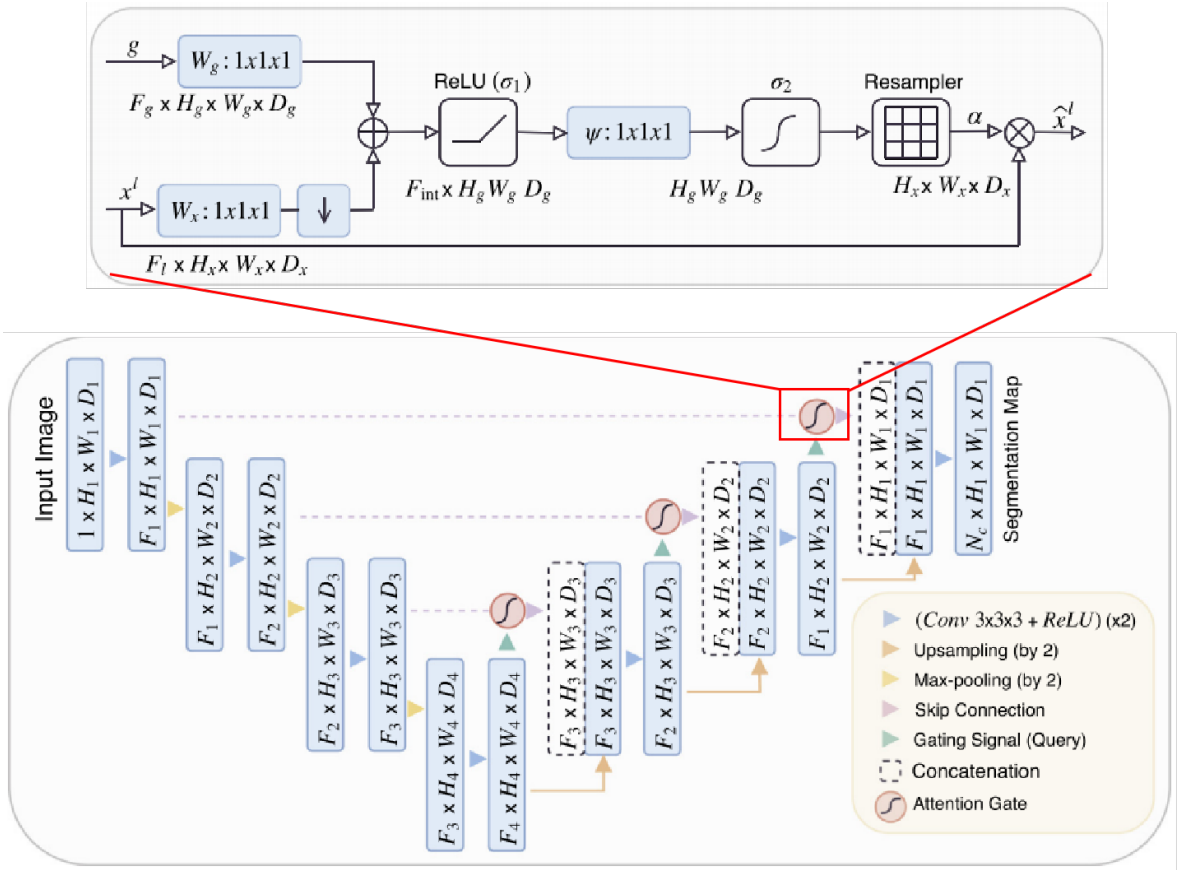

Figure 2.28: Architecture of Attention U-Net (Schlemper et al., 2019).

irrelevant regions) and the shallow features (has more irrelevant regions) from the encoder as the reference vector. The attention gates of different scales were multiplied with their corresponding shallow features to disambiguate task-irrelevant information before the concatenation of the shallow and deep features, which encourages the network to focus more on the target regions during both forward and backward passes and thus obtained segmentation results with higher accuracy.

Therefore, a typical CNN model that normally contains convolutional layers-batch normalization-activation function (a.k.a. nonlinearity layer) block and pooling layer (a.k.a. down-sampling layer) is established. For classification models, the fully connected layer is indispensable, while the up-sampling layer is necessary for segmentation models. In particular, the backpropagation and optimization algorithms are the foundation for the ability of CNNs to learn target features, and their emergence has played an indelible role in the development and application of CNNs. Their detailed introductions are as follows:

Backpropagation Algorithm

Backpropagation, short for "error backpropagation," is used in conjunction with optimization methods to train deep learning models. The method computes the gradients of the loss function for all the weights in the network. These gradients are then fed back to the optimization method and used to update the network weights to minimize the loss function.

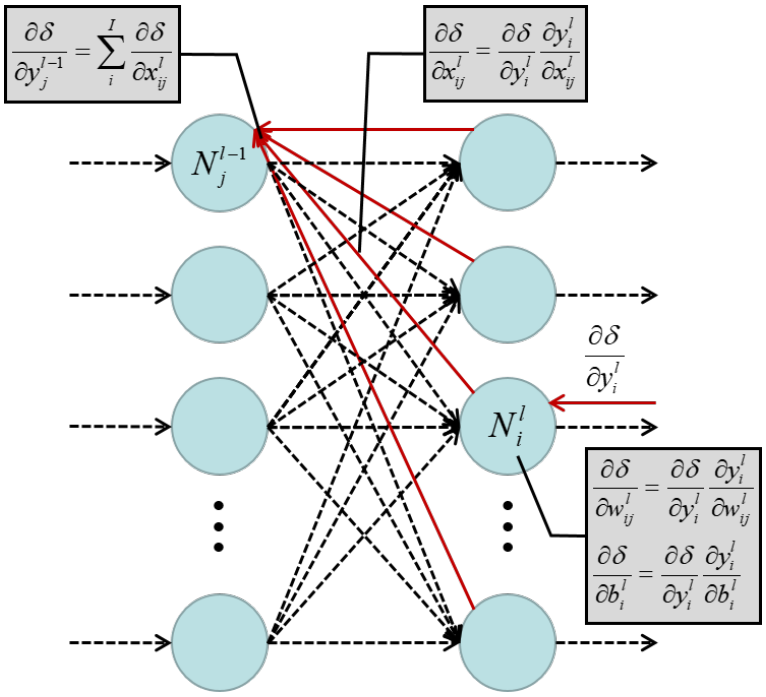

Figure 2.29: Illustration of the backpropagation algorithm.

Fig. 2.29 gives an illustration of the backpropagation algorithm. Provided the prediction error between the network output and the ground truth is δ, then the δ derivative in terms of the output y_i^l of the ith neuron N_i^l in the lth layer is calculated

as $\frac{\partial \delta}{\partial y_i^l}$. The error derivative in terms of the parameters between the neuron N_i^l and the jth neuron N_j^{l-1} in the $l-1$th layer is calculated as:

$$\frac{\partial \delta}{\partial w_{ij}^l} = \frac{\partial \delta}{\partial y_i^l} \frac{\partial y_i^l}{\partial w_{ij}^l} \tag{2.1}$$

$$\frac{\partial \delta}{\partial b_i^l} = \frac{\partial \delta}{\partial y_i^l} \frac{\partial y_i^l}{\partial b_i^l} \tag{2.2}$$

where w_{ij}^l and b_i^l are the weight and bias between neurons N_i^l and N_j^{l-1}, respectively. For the error derivative in terms of the output y_j^{l-1} of neuron N_j^{l-1}, it is calculated as the summation of the feedback from all the neurons in the lth layer connected to the neuron N_j^{l-1}, which is expressed as:

$$\frac{\partial \delta}{\partial x_{ij}^l} = \frac{\partial \delta}{\partial y_i^l} \frac{\partial y_i^l}{\partial x_{ij}^l} \tag{2.3}$$

$$\frac{\partial \delta}{\partial y_j^{l-1}} = \sum_i^I \frac{\partial \delta}{\partial x_{ij}^l} \tag{2.4}$$

where $x_{ij}^l = y_j^{l-1}$ is the input to neuron N_i^l from neuron N_j^{l-1}, and I is the total neuron number in the lth layer. With the recursive application of the chain derivation rule to the network from the output to the input, the error derivative in terms of each layer's parameters is calculated.

Optimization Algorithms

Optimization algorithms are used to guide the update of network parameters towards appropriate values with a certain stride based on the gradients calculated according to the loss function and backpropagation during the training, which forces the loss function (objective function) value to approach the global minimum. So far, researchers have developed many optimization algorithms, such as stochastic gradient descent (SGD) (Bottou, 2012), SGD with momentum (SGD-M) (Liu et al., 2020b), SGD with nesterov momentum (SGD-NM) (Ruder, 2016), adaptive gradient (Adagrad) (Wilson et al., 2017), Adadelta (Zeiler, 2012), RMSprop (Zou et al., 2019), adaptive moment estimation (Adam) (Kingma and Ba, 2014), Adamax (Kingma and Ba, 2014), and Nadam (Dozat, 2016). Among all optimization algorithms, SGD, SGD-M, and Adam are the most favored model optimization algorithms adopted by researchers.

SGD algorithm calculates the mini-batch gradients of each iteration and update the network parameters with the average of the obtained gradients and a manual set learning rate:

$$g_t = \frac{1}{m} \sum_i^m \nabla_{\theta_{t-1}} f_i\left(\theta_{t-1}\right) \tag{2.5}$$

$$\theta_t = \theta_{t-1} - \eta g_t \tag{2.6}$$

where g_t and $f_i(\theta_{t-1})$ are the average of the mini-batch gradients at time t and the loss computation of ith sample in the mini-batch at time $t-1$, respectively, and m is the mini-batch size. η is the learning rate. θ_t and θ_{t-1} are the updated network parameters at time t and network parameters at time $t-1$ respectively. The fluctuations caused by SGD favor the direction of optimization to jump from the current local minima to another better local minima, such that for non-convex functions, it eventually converges to better local minima or even global minima. However, when a local optimum point or saddle point is reached, the gradient is 0, and the parameter update cannot continue. It oscillates along the steep direction while progressing slowly along the gentle dimension, making it difficult to converge quickly.

SGD-M algorithm adds a first-order momentum to the SGD:

$$g_t = \beta g_{t-1} + (1-\beta)\frac{1}{m}\sum_{i}^{m}\left(\nabla_{\theta_{t-1}} f_i(\theta_{t-1})\right) \tag{2.7}$$

$$\theta_t = \theta_{t-1} - \eta g_t \tag{2.8}$$

where β is a momentum hyper-parameter, g_{t-1} is the average of the mini-batch gradients at time $t-1$. Because of the addition of momentum, SGD-M alleviates the problem that SGD cannot be continuously updated when the gradient of the local optimum is 0 and the problem that the oscillation amplitude is too large. Yet, these problems are not completely solved. When the local gully is deep while the momentum is used up, it will still be trapped in the local optimum and oscillate back and forth.

Adam algorithm is built on the SGD-M and utilizes the first-order and second-order moment estimations of the gradient to adaptively and dynamically adjust the learning rate of each parameter:

$$h_t = \alpha h_{t-1} + (1-\alpha)\frac{1}{m}\sum_{i}^{m}\left(\nabla_{\theta_{t-1}} f_i(\theta_{t-1})\right) \tag{2.9}$$

$$v_t = \mu v_{t-1} + (1-\mu)\left(\frac{1}{m}\sum_{i}^{m}\left(\nabla_{\theta_{t-1}} f_i(\theta_{t-1})\right)\right)^2 \tag{2.10}$$

$$\hat{h}_t = \frac{h_t}{1-\alpha^t} \tag{2.11}$$

$$\hat{v}_t = \frac{v_t}{1-\mu^t} \tag{2.12}$$

$$\theta_t = \theta_{t-1} - \eta\frac{\hat{h}_t}{\sqrt{\hat{v}_t}+\varepsilon} \tag{2.13}$$

where h_t and v_t are the first-order and second-order moment estimations of the gradient, respectively. $\hat{h}_t$ and $\hat{v}_t$ are their corresponding calibrations. α and μ are the first-order and second-order momentum hyper-parameters, respectively. First-order and second-order moment estimations effectively control the learning rate and gradient descent direction. They also prevent the oscillation of the gradient and the stationary at the saddle point.

2.3.2 Weakly-Supervised Learning

Supervised learning is the most commonly used learning method for deep learning models, which requires training datasets with full labels. In particular, for segmentation or localization tasks, the pixel-wise label indicating the object's location and contour (target region) information is necessary for supervised learning. However, in many cases, obtaining strong labels at the pixel level is very expensive and time-consuming, while relatively weak labels at the image level are much easier to produce. This fact has driven the birth and flourishing development of another learning method, i.e., weakly-supervised learning. The principle of weakly-supervised learning for segmentation or localization (a.k.a. weakly-supervised segmentation or localization) is that a deep learning model is expected to provide pixel-wise annotations based on the training with only image-wise labels.

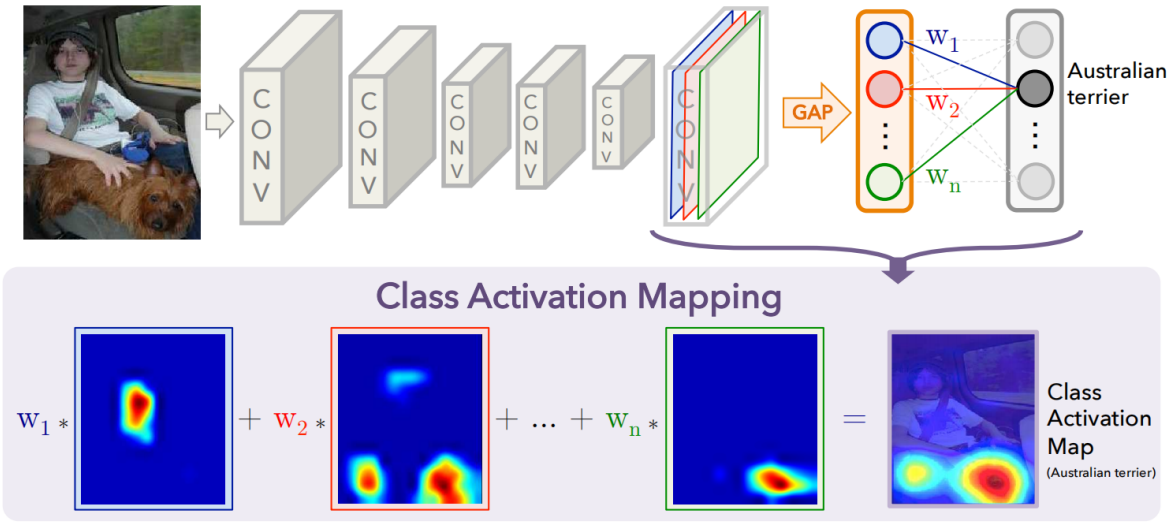

Figure 2.30: Architecture of CAM (Zhou et al., 2016).

In 2016, Zhou et al. proposed a general module named class activation mapping (CAM) along with global average pooling to achieve weakly-supervised object localization, which obtained a top-5 error of 37.1% on ILSVRC 2014 (Russakovsky et al., 2015). Fig 2.30 shows the architecture of the CAM. The CAM on top of a CNN enables the visualization of the predicted class scores on the input image and highlights the discriminative regions of the object, which reveals the implicit attention of the CNN on the image.

In 2017, Durand et al. proposed a model called WILDCAT to achieve image classification, point-wise localization, and segmentation. Fig. 2.31 provides the architecture of the WILDCAT. It was built on FCN (Long et al., 2015) and ResNet-101 (He et al., 2016b) and is a typical weakly-supervised segmentation and localization method. WILDCAT used a multi-map WSL transfer layer that learns multiple class-related modalities to generate M feature maps of each class, which were then fed into a spatial aggregation module to produce class-corresponded probability, localization information, and segmentation result.

The same year, Selvaraju et al. proposed the Grad-CAM (Selvaraju et al., 2017a) to produce visual explanations of many CNN models, which makes them more general

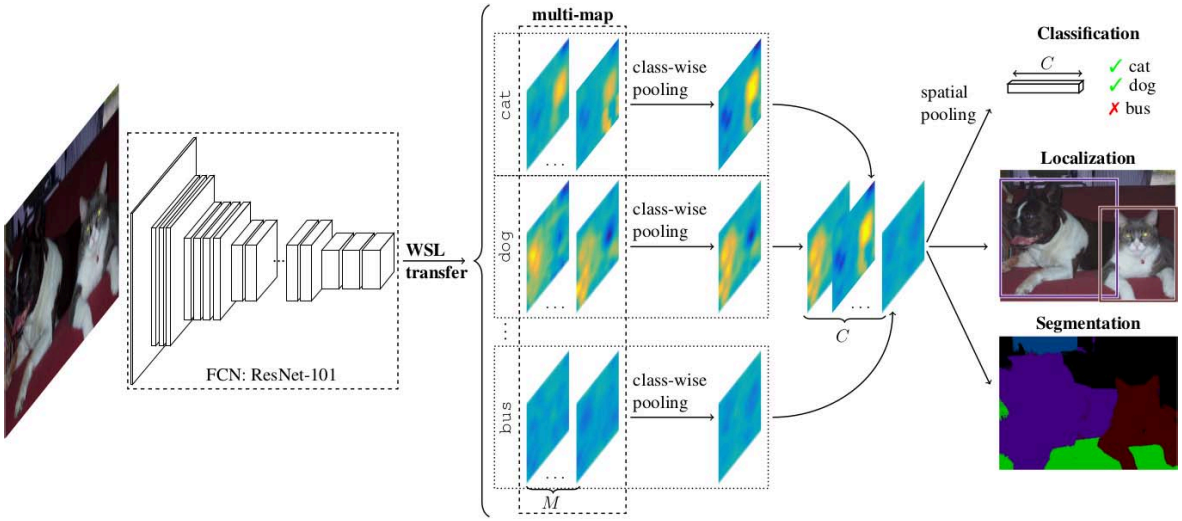

Figure 2.31: Architecture of WILDCAT (Durand et al., 2017).

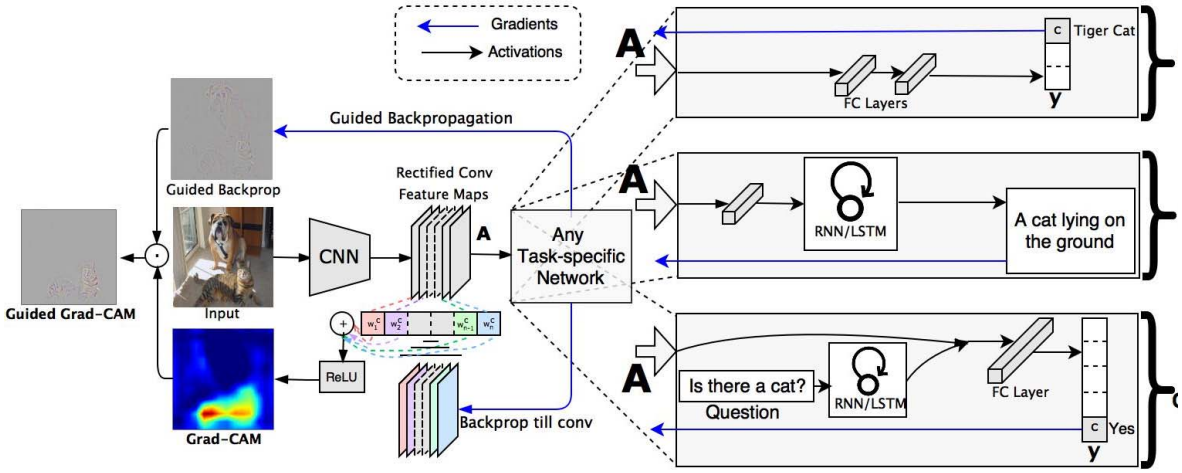

Figure 2.32: Architecture of Grad-CAM (Selvaraju et al., 2017a).

in bringing transparency. Fig. 2.32 gives the architecture of the Grad-CAM. It used the gradients of a specific target flowing into the final convolutional layer to generate a coarse localization map, which was then combined with existing fine-grained visualizations to underline the key regions in the image that support the prediction. The Grad-CAM was then improved to Grad-CAM++ by Chattopadhay et al. in 2018 (Chattopadhay et al., 2018), which is giving in Fig. 2.33. The main improvement is that they used more generalized weights to produce the visual explanation for the class label with high probability. The weights are calculated based on a weighted combination of the feature maps positive partial derivatives in terms of a specific class score in the last convolutional layer.

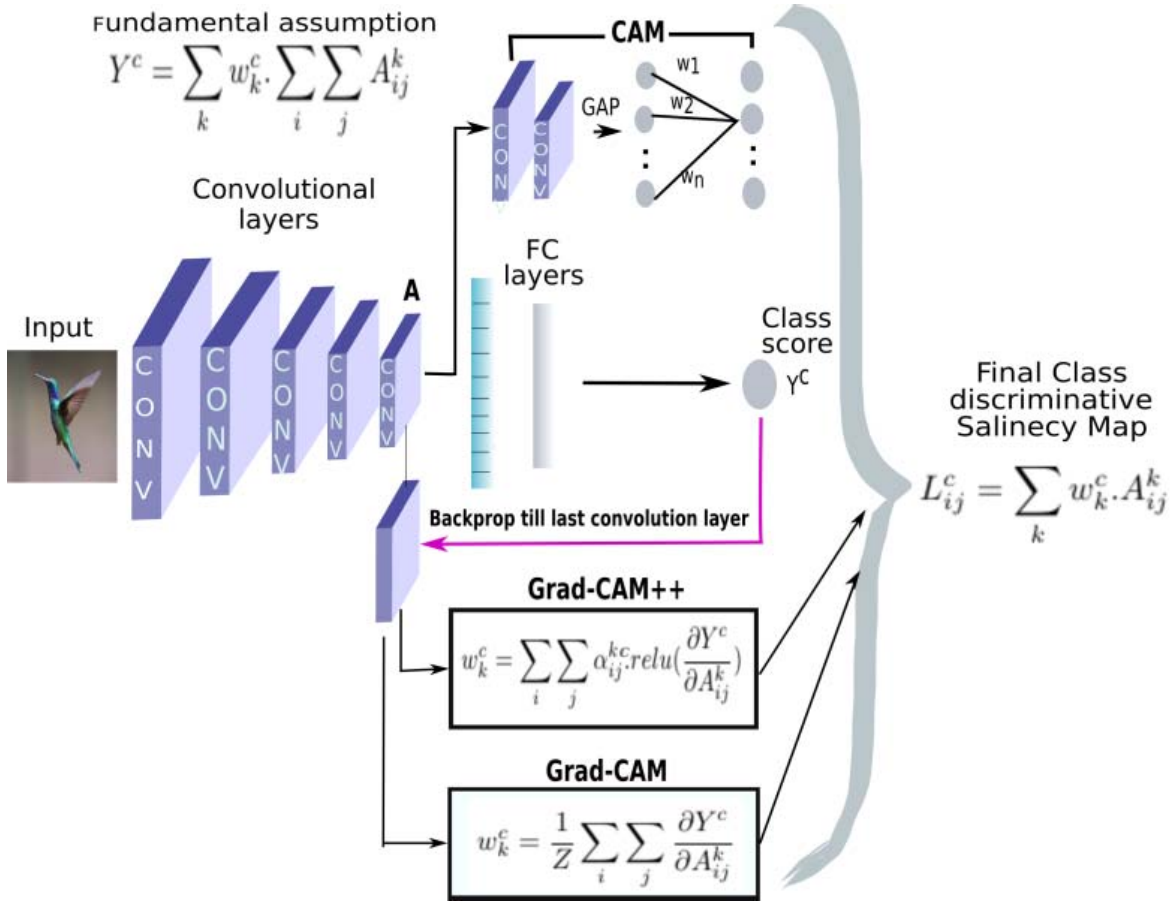

Figure 2.33: Architecture of Grad-CAM++ (Chattopadhay et al., 2018).

In 2018, Zhang ea al. proposed an ACoL model Zhang et al. (2018a) based on the adversarial mechanism to achieve the integral localization of objects of interest through weak supervision. Fig. 2.34 shows the architecture of the ACoL. They designed two parallel classifiers on top of a feature extraction backbone coming from VGG16 (Simonyan and Zisserman, 2014b) or GoogLeNet (Szegedy et al., 2015b) to obtain the object localization and classification results, where one classifier first localizes part of discriminative object regions and then drives another classifier to discover extra and complementary object regions via erasing its recognized parts from the feature maps.

A-MIL (Ilse et al., 2018a; Patil et al., 2019) is a special weakly-supervised localization method. Fig. 2.35 gives the architecture of the A-MIL. Unlike general methods that

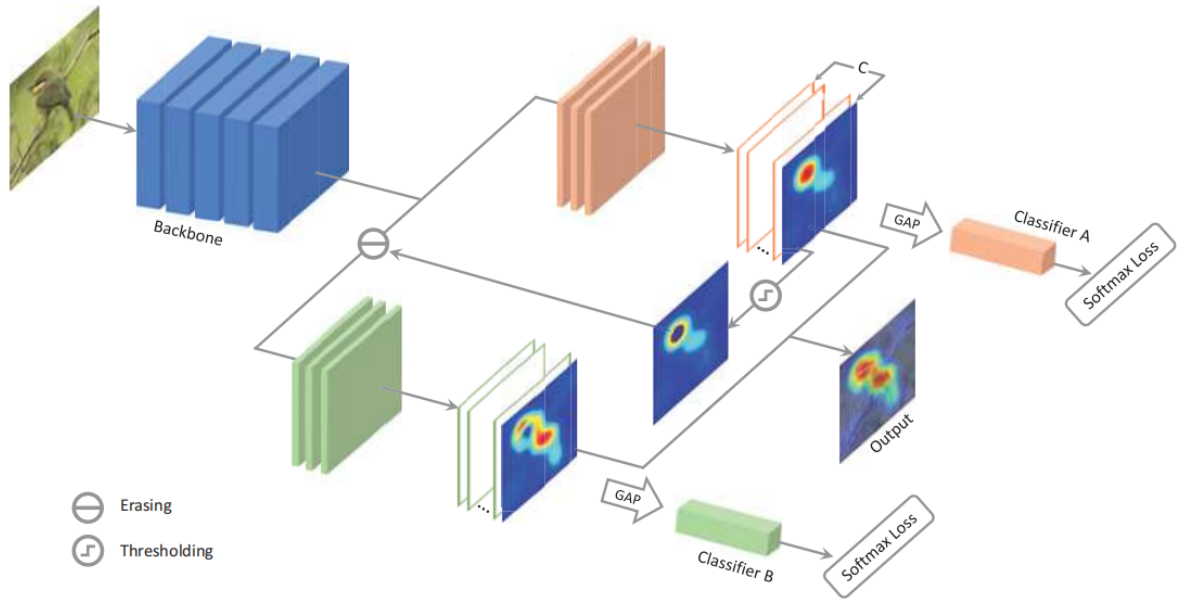

Figure 2.34: Architecture of ACoL (Zhang et al., 2018a).

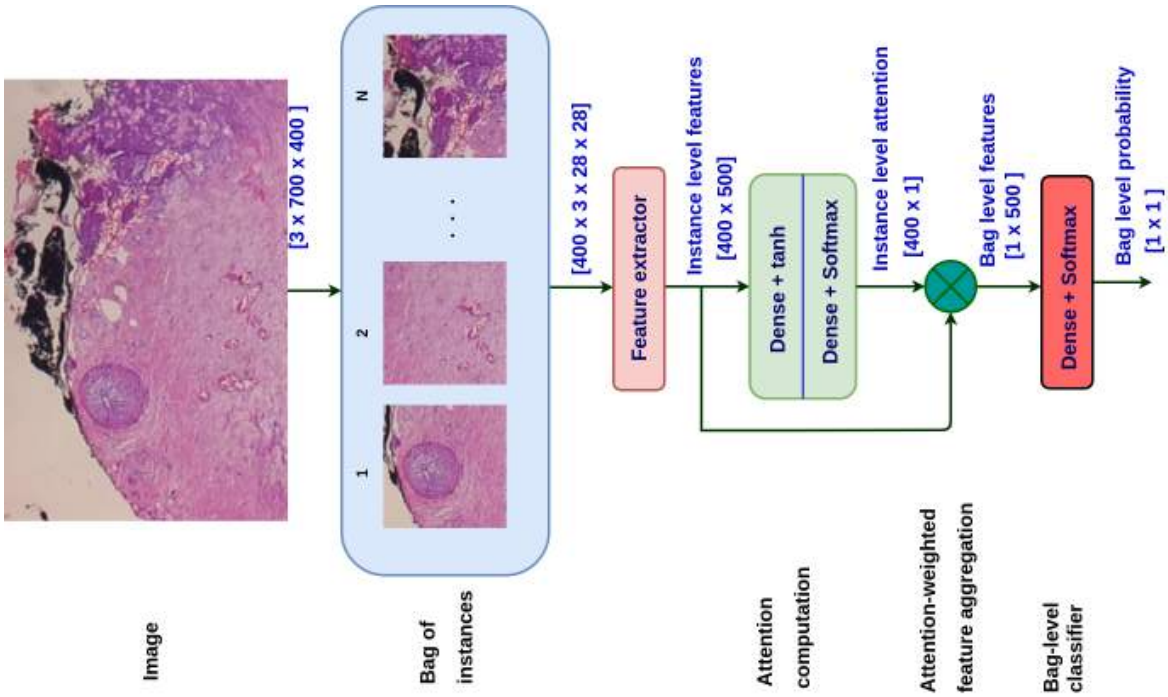

Figure 2.35: Architecture of Attention MIL (Ilse et al., 2018a; Patil et al., 2019).

assign values to each pixel, It first divided the input image into a bag of small patches (up to hundreds and thousands), which are treated as pixels and eventually assigned category values, e.g., 0 or 1, by an attention-based MIL pooling module learned during the training with only bag-level (i.e., image-level) labels. These assigned patches are then pieced together to form a heat map or binary mask showing the localization of the target.

In 2019, Huang et al. Huang and Chung (2019) proposed a CELNet model (as shown in Fig. 2.36) with multi-branch attention modules and a deep supervision mechanism to address the difficulty of classifying histopathological breast cancer images. After obtaining the classification result, they used the combination (CELM) of Grad-CAM (Selvaraju et al., 2017a) and guided back-propagation (Springenberg et al., 2014) to generate the evidence localization to explain the decision.

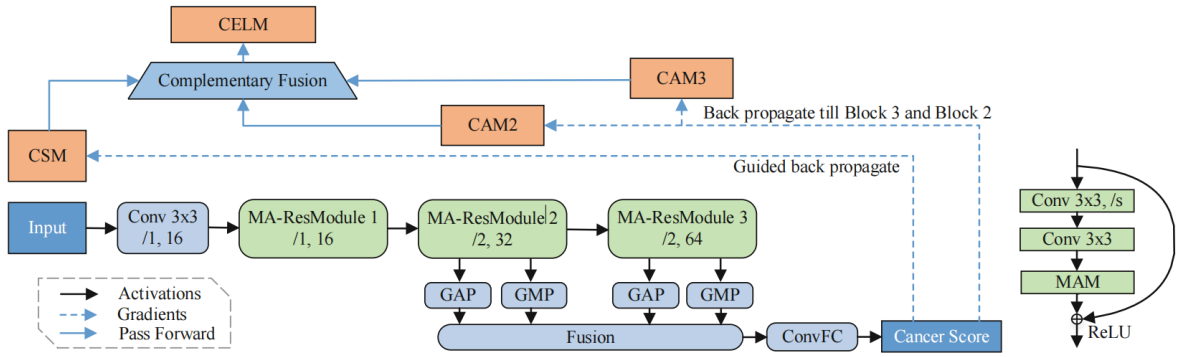

Figure 2.36: Architecture of CELNet-CELM (Huang and Chung, 2019).

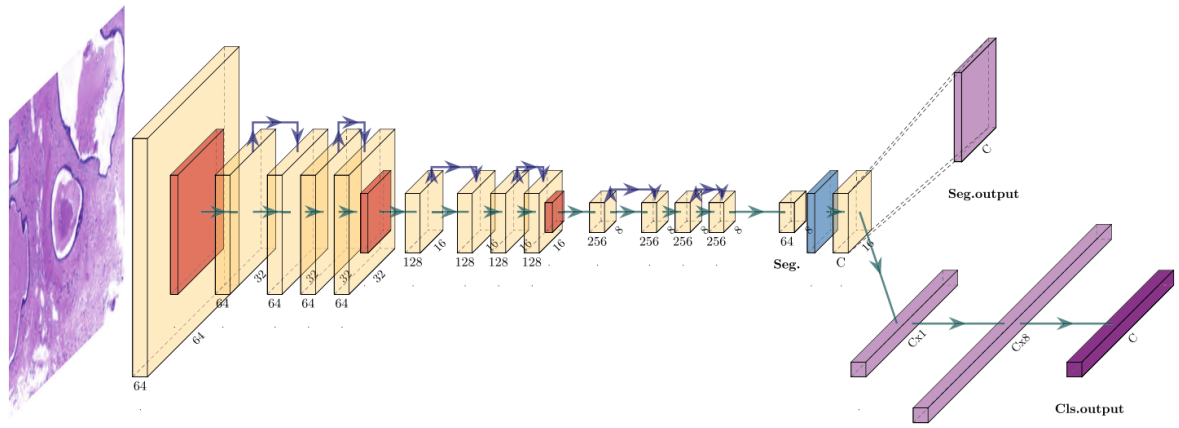

Figure 2.37: Architecture of the model proposed by Ciga et al. Ciga and Martel (2021).

More recently in 2021, Ciga et al. Ciga and Martel (2021) utilized the third layer (out of a total of four layers) output of the ResNet-18 architecture (He et al., 2016b) for encoding input images. Fig. 2.37 provides the architecture of their model. They then developed a simple spatial up-sampling module on top of this feature encoder to generate a low-resolution segmentation output (i.e., segmentation result before spatial interpolation), which was fed into a fully convolutional classifier to obtain the classification result.

2.4 Summary

In this chapter, we describe the research objects and the deep learning method relevant to this thesis:

- First, this chapter provides a more detailed introduction to breast cancer diagnosis, including types of breast cancer, imaging techniques for breast disease diagnosis, and two public datasets for studying computer-aided diagnostic system algorithms used to help doctors or radiologists give more accurate and objective classification decisions for clinical histopathological images.
- Secondly, this chapter provides an introduction to display panel defect detection, including the process of display panel manufacturing, algorithms of automated optical inspection systems for online defect detection, and the dataset for studying automated optical inspection system algorithms that are applied to relieve the pressure of manual visual inspection by shop floor staff and to achieve highly efficient online defect detection.
- Finally, this chapter describes two techniques closely related to this thesis. One is convolutional neural networks by introducing their origin and development as well as two important algorithms that underpin their implementation. The other is weakly-supervised learning as well as its definition and development.

Chapter 3

Configurable Convolutional Neural Networks

3.1 Introduction 78
3.2 Proposed Configurable Network 79
3.3 Breast Cancer Classification with Visual Explanation via the FEM-DMG-Classifier Configuration of ConfigNet 83
3.3.1 Introduction 83
3.3.2 Methodology 84
3.3.3 Experimental Settings 88
3.3.4 Results and Discussion 89
3.3.5 Conclusion 92
3.4 Breast Tumor Segmentation via the Encoder-Decoder Configuration of Our ConfigNet 93
3.4.1 Methodology 93
3.4.2 Experimental Settings 97
3.4.3 Results and Discussion 98
3.4.4 Conclusion 102
3.5 Summary 102

3.1 Introduction

In the last two decades, deep-learning methods, particularly convolutional neural networks (CNNs), have been the hottest research point and been widely utilized in many fields of image processing, such as target detection (Han et al., 2022; Wang et al., 2022), object classification (Huang et al., 2017; Ma et al., 2018), and semantic segmentation (Lin et al., 2022; Zhou et al., 2022) due to their automatic feature learning and superior feature representation abilities. Specifically, VGG (Simonyan and Zisserman, 2014a), ResNet (He et al., 2016a), Inception models(Szegedy et al., 2015a, 2016, 2017), and DenseNet (Huang et al., 2017) were developed to achieve nature images (from ImageNet dataset) classification; RCNN and its series (Girshick et al., 2014; He et al., 2017; Ren et al., 2015; He et al., 2017), YOLO and its series (Redmon et al., 2016; Redmon and Farhadi, 2017, 2018; Bochkovskiy et al., 2020; Wang et al., 2022), and BoxeR (Nguyen et al., 2022) were proposed for target detection; FCN (Long et al., 2015), U-Net and its series (Ronneberger et al., 2015; Alom et al., 2018; Schlemper et al., 2019; Thomas et al., 2020; Zunair and Hamza, 2021; Lin et al., 2022), SegNet (Badrinarayanan et al., 2017), and HSSN (Li et al., 2022) were developed for semantic segmentation of different objects.

Inspired by the considerable breakthroughs and success of the deep learning methods in classification, localization, and segmentation tasks in medical image analysis (Ronneberger et al., 2015; Carneiro et al., 2017; Schlemper et al., 2019; Xu et al., 2019a; Ibtehaz and Rahman, 2020) and defect detection (Zou et al., 2018; Lian et al., 2019; Liu et al., 2019; Dong et al., 2019; Song et al., 2020; Huang et al., 2021b), we attempt to leverage the superior performance of CNNs to achieve our goals. It leads to the two main objectives of this thesis mentioned in Chapter 1. First of all, we attempt to develop a deep learning model to achieve effective and accurate classification of histopathological breast cancer images while providing visual explanations (i.e., introducing explainability or interpretability and transparency to CNN that has the "black box" nature), which is of great importance for the safety, ethics, trustworthy and reliability in clinical diagnosis and for the deployment of deep learning-based CAD systems in real-world clinical settings. Secondly, we attempt to develop a high-efficient and accurate deep learning model for automated optical inspection (AOI) systems to solve the problem of real-time defect detection of display panels on factory assembly lines, which plays a vital role in the improvement of the yield rate.

Nevertheless, the current deep learning network architecture for a specific target task is generally fixed and not suitable for other tasks. For instance, an effective CNN model for a segmentation or localization task should provide pixel-level predictions, which requires the model to have an encoder-decoder structure to generate end-to-end results where the decoder requires up-sampling processes. Classification tasks require a CNN model to extract the most discriminative information of the target region to accurately identify different categories, which makes it indispensable for the CNN model to have a classifier at the end to produce categories probabilities, and the probabilities are image-level predictions. Furthermore, different investigation objects of the same task are often very variable, such as the semantic segmentation of breast tumors (see Fig. 3.1-top) and display panel defects (see Fig. 3.1-bottom). Since there are many differences in their image textures, the deep learning models have diverse requirements

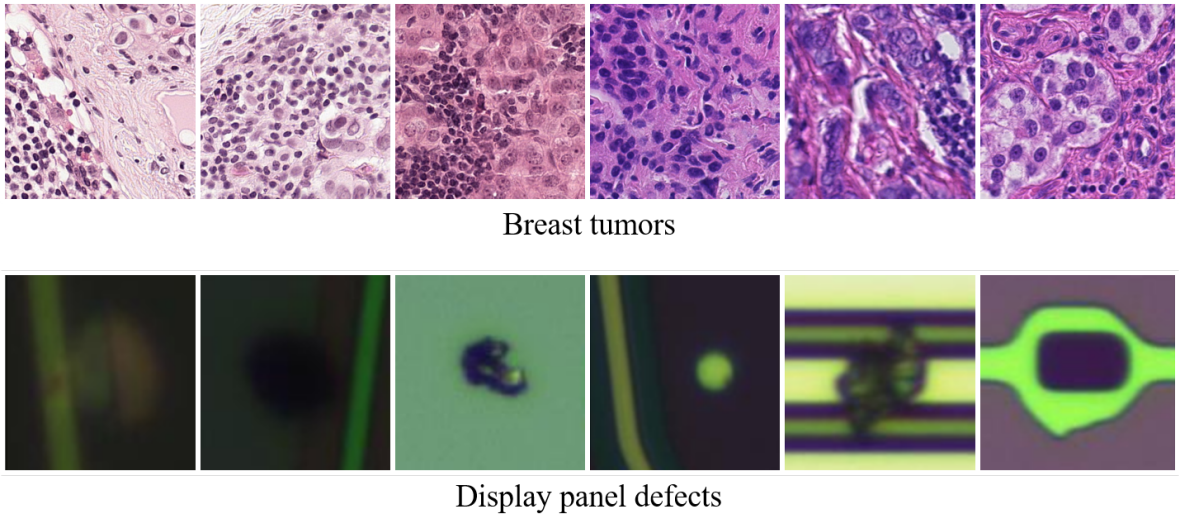

Figure 3.1: Typical examples of breast tumor images (top row) and display panel defect images (bottom row).

for the feature extraction of these two objects (i.e., breast tumors and display panel defects). In addition, the former is more favored for accuracy, while the latter is more favored for detection speed. Therefore, we developed a configurable convolutional neural network (ConfigNet) capable of transforming into different configurations to adapt to multiple tasks and objects, which is compatible with both explainable classifications of histopathological breast cancer images and online defect detection of display panels.

3.2 Proposed Configurable Network

The overall architecture of our ConfigNet (i.e., configurable network) is depicted in Fig. 3.2. It is mainly composed of two backbone branches that corresponds to two configurations adapted to different tasks (e.g., classification and segmentation).

One is FEM-DMG-classifier configuration that is applied to achieve explainable image classification, where FEM (i.e., feature extraction module) is aimed to extract target features, DMG (i.e., decision map generator) is dedicated to generating decision maps (a.k.a. categories confidence maps) for explanations, and classifier is devoted to making decisions. As observed with more details in Fig. 3.3, it is only trained with image-level ground truth (e.g., normal or tumor) and capable of providing categories probabilities (classification) and pixel-level predictions (explanation). In order to leverage the superior performance of existing deep learning networks, the FEM is configured via the transfer learning strategy, i.e., using the shallow convolutional layers of existing CNN models (such as VGG (Simonyan and Zisserman, 2014a) and ResNet (He et al., 2016a)) as our FEM. The DMG is a module that copes with the extracted abstract features from FEM and generates maps with their channels corresponding directly to categories, thus providing evidence for the subsequent classification. Two configurations of the DMG were developed in this thesis. One is the basic DMG with a simple structure

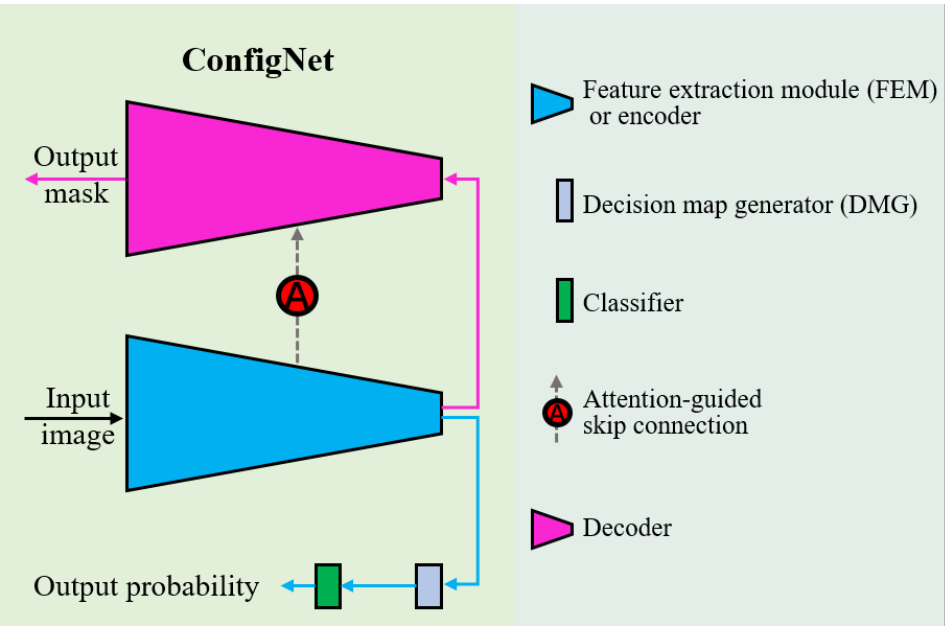

Figure 3.2: Illustration of ConfigNet

(detailed in Section 3.3), and another is the improved DMG with multi-size receptive fields (detailed in Chapter 4). The classifier is used to model the decision maps of the DMG into probabilities with a one-to-one correspondence between map channels and categories. Global average pooling is a classic classifier for this purpose. We proposed a weighted average pooling (WAP) classifier for the FEM-DMG-classifier configuration of our ConfigNet to achieve better performance, which will be detailed in Section 3.3.

The other one is the encoder-decoder configuration that is applied to achieve target segmentation and localization, where the encoder (a.k.a. FEM) is a module that encodes the input image into maps with hundreds and thousands of channels containing different features of the image while the decoder with attention-guided skip connection is to transform these feature maps into masks (e.g., binary mask) with regions (each contains pixels of the same value) indicating locations and contours of targets. More details of the encoder-decoder configuration are shown in Fig. 3.4. It is trained with pixel-level ground truth in an end-to-end way. Similar to the FEM-DMG-classifier configuration, the encoder is constructed based on the transfer learning and fine-tuning of existing CNNs with deep convolutional layers having prominent performance in feature extraction. The decoder is built according to the encoder and has an approximately symmetrical structure, which can be configured to favor different purposes. The attention-guided skip connection (a.k.a. feature fusion module, abbreviated as FFM) between the encoder and decoder (i.e., shallow and deep layers) is developed to encourage the network to focus more on the target regions while discrediting the responses of the background. We have proposed two configurations of the FFM for the encoder-decoder configuration of our ConfigNet. One is an efficiency-favored element-wise feature fusion module (EFFM) built on the attention gate of Attention U-Net (Schlemper et al., 2019) and is used for online defect detection (detailed in Chapter 5). Another is an accuracy-favored spatial and channel attention-guided feature fusion module (SCAFFM) that is designed for medical image segmentation (detailed in

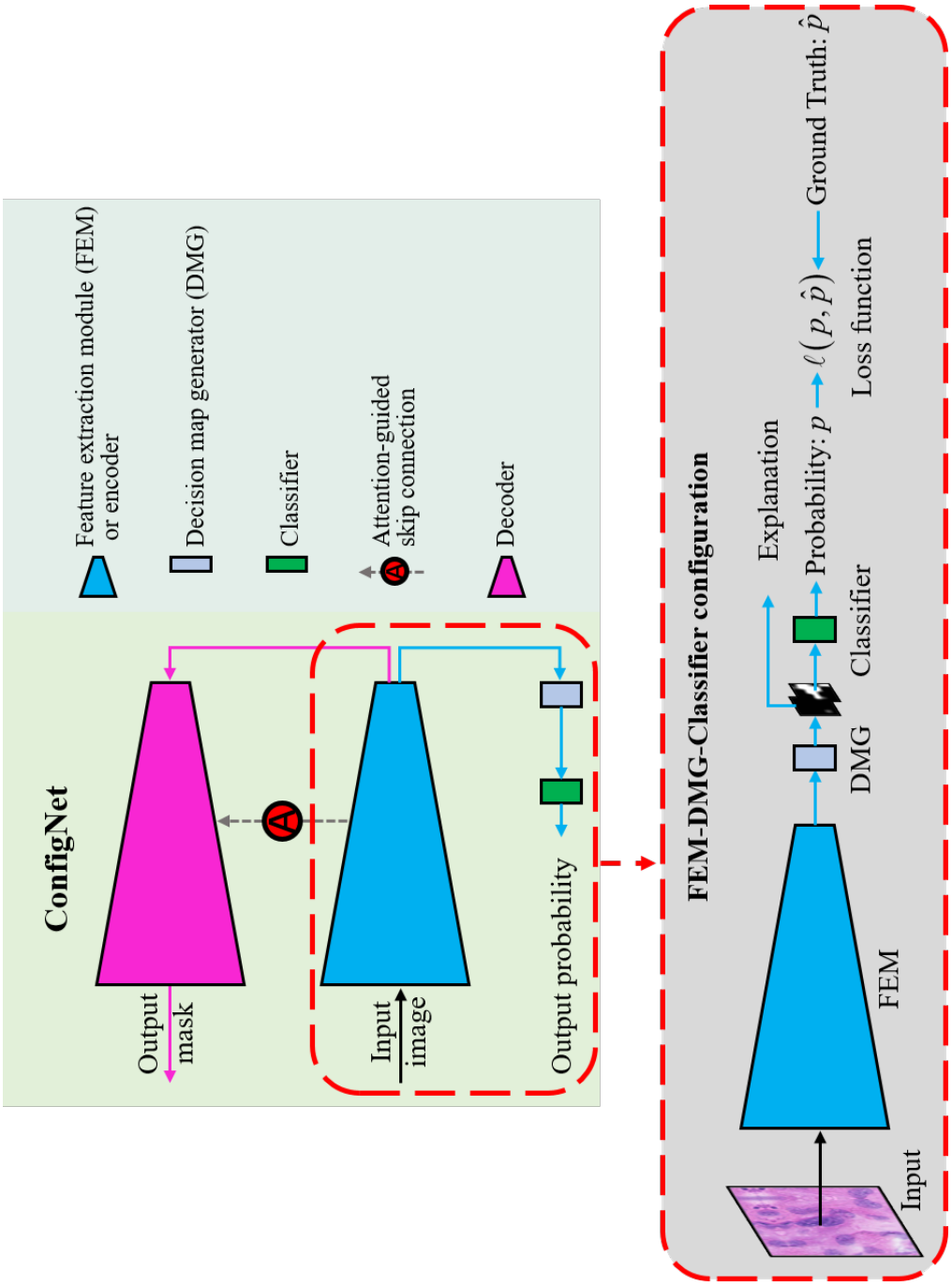

Figure 3.3: Illustration of the FEM-DMG-classifier configuration.

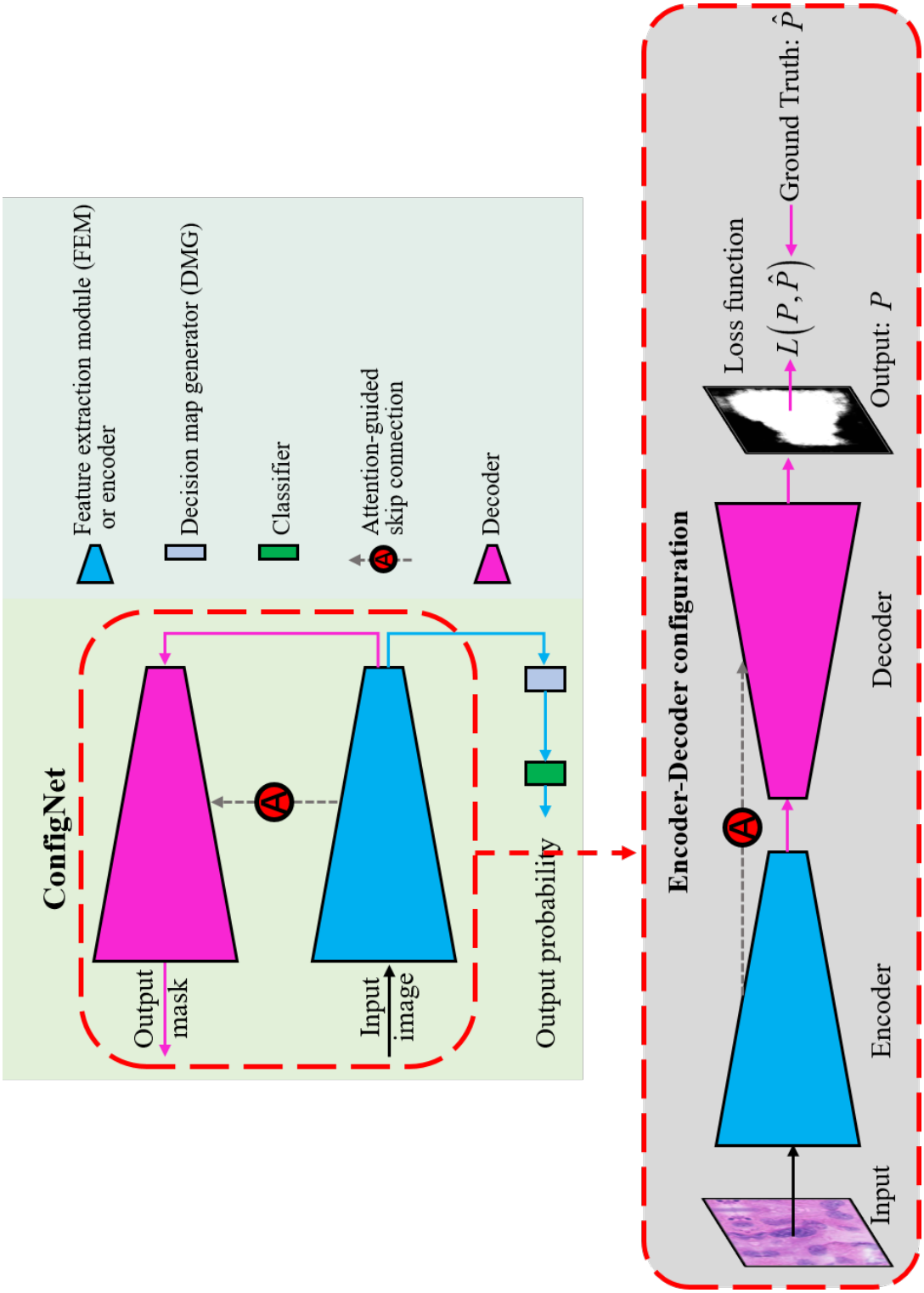

Figure 3.4: Illustration of the encoder-decoder configuration.

Section 3.4).

To evaluate the effectiveness of our ConfigNet in multiple tasks, such as explainable classification and tumor segmentation of breast cancer, we performed experiments on two configurations, respectively, on two publicly available breast cancer datasets. The first one is the FEM-DMG-classifier configuration based on VGG, basic DMG, and WAP classifier; the details are given in Section 3.3. The second one is the encoder-decoder configuration based on VGG and SCAFFM, which is described in detail in Section 3.4.

3.3 Breast Cancer Classification with Visual Explanation via the FEM-DMG-Classifier Configuration of ConfigNet

3.3.1 Introduction

Generally, clinical whole-slide histopathological images have a large size (billions of pixels), which makes it highly time-consuming and labor-intensive for pathologists to give local annotations and thus causes a lack of available data. It hinders the performance of localization and segmentation deep learning models in clinical use. The large size also undoubtedly burdens the network's forward propagation and brings computational infeasibility for deep learning-based methods.

On the other hand, an effective deep learning convolutional neural network for classification must have two main components: a feature extractor with enough depth and width to extract global and discriminative contextual information and a classifier with the ability to translate abstract features into classification scores that decide the result. In general, the fully connected layer (also known as linear layer) followed by a nonlinear activation function is used as the classifier of a network due to its property of connecting all the nodes of the previous layer to all the nodes of the current layer, which has an advantage in integrating the extracted features and enhance the model learning ability. However, its non-linearity and complex node connections hinder the transparency and comprehensibility of the model. A global average pooling operator (Gao et al., 2014) was proposed to replace the fully connected layer to improve the classification interpretability by compulsively transforming the abstract feature maps into categories confidence maps that correspond to categories through learning. Yet, it also places a higher demand on the network's feature extractor compared to the one with fully connected layer. While the increase of the down-sampling process (such as max-pooling) and feature extractor depth or width enhances the feature extractor performance in identifying global and discriminative information relevant to the target, it either reduces the image resolution that unavoidably leads to the loss of some important information helpful for improving classification accuracy or increases the network's computational complexity, which can be a disaster for inputs of large size.

Multiple instance learning (MIL) (Dietterich et al., 1997; Zhou, 2004; Quellec et al., 2017) is a special deep learning strategy that provides an elegant framework to deal with the above issue. According to the MIL principle, the inputs of a model are bags (sets) of instances with only bag-level labels, while the annotation for each instance is

unnecessary. In this way, a large-scale image can be cropped into many small instances and fed into a deep learning model as a bag, which enables the model to process the image more efficiently. Therefore, MIL has been widely used to develop effective deep learning methods to achieve the classification (Chikontwe et al., 2020; Zhao et al., 2020; Shao et al., 2021), localization (Li et al., 2021a; Rony et al., 2019), and segmentation (Xu et al., 2019b; Lerousseau et al., 2020) of histopathological images. For example, Jia et al. (Jia et al., 2017) proposed a MIL framework based on deep weak supervision, end-to-end learning, and FCNs to segment cancerous regions in histopathological images with large scales. The performance of their model in terms of F1-measure for the segmentation of colon cancer is 83.6%. Sudharshan et al. (Sudharshan et al., 2019) proposed a weakly supervised learning framework for the computer-aided diagnosis (CAD) of breast cancer patients. Their investigation showed that MIL can leverage the classification and analysis of histopathological images to improve computer-aided diagnosis. Vu et al. (Vu et al., 2020) proposed a novel symmetric MIL framework that associated each instance in a bag with its attribute being either negative, positive, or irrelevant to give image-level and region-level annotations for histopathological images. They presented experiments on 7 real-world datasets and obtained competitive results. Sharma et al. (Sharma et al., 2021) proposed an end-to-end MIL framework for the classification of whole slide images (WSI). In their work, the WSI was divided into clusters of patches, an adaptive attention mechanism was used to give a slide-level prediction, and KL-divergence loss was adopted to optimize the model. Experiments on the Camelyon16 dataset showed that their framework was able to classify breast cancer at slide level with an accuracy of 91.12%.

Inspired by the great effectiveness and success of MIL methods in the above-mentioned works, we introduced the MIL strategy into the FEM-DMG-classifier configuration of our ConfigNet to develop an easy-to-implement model (multi-instance classification network, abbreviated as MICNet) that solves the problem of explainable classification of histopathological breast cancer images.

3.3.2 Methodology

The overview of our MICNet including input image preprocessing is shown in Fig. 3.5. We used the image-level binary label to train the model to make it able to produce the classification decision while providing a logical visual explanation that indicates the presence (and its localization) or absence of the tumor.

Multiple Instance Learning

MIL (i.e., multiple instance learning) is a weakly-supervised learning strategy that aims to find a model to provide a single label for a bag (or set) of instances. Specifically, let $X_n = \{x_1, x_2, ...x_M\}$ be a bag defined as a set of feature vectors, this bag of instances (i.e. feature vectors) has only one bag-level label Y_n. In standard MIL, Y_n is assumed as negative only if all the instances are negative and is positive if one or more instances are positive, i.e.,

$$Y_n = \begin{cases} 1, & \text{if } \exists x_i \in X_n : x_i = 1 \\ 0, & \text{otherwise} \end{cases} \tag{3.1}$$

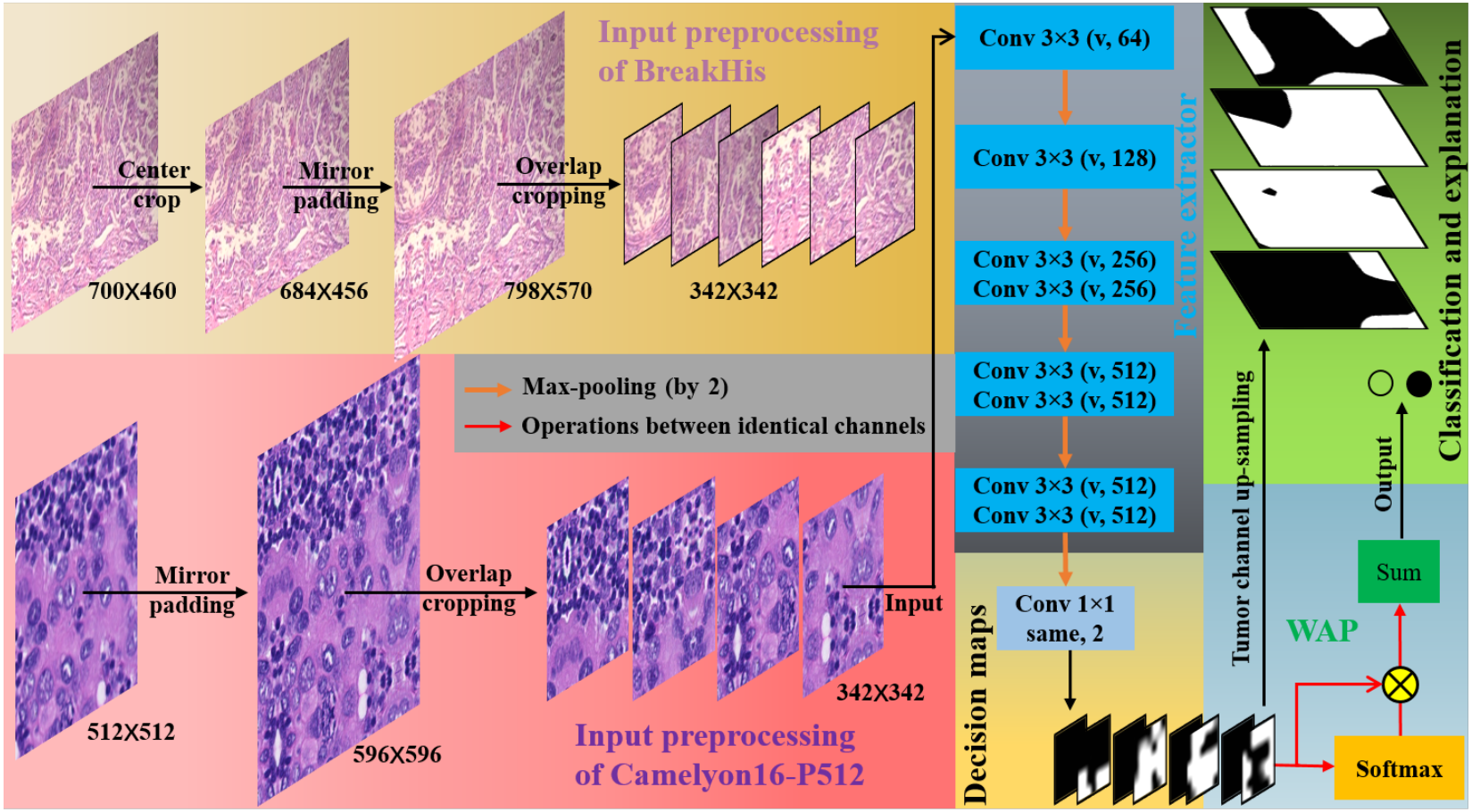

Figure 3.5: An illustration of MICNet architecture.

which can be re-formulated in a compact by using the maximum operator:

$$Y_n = \max\{X_n\} \tag{3.2}$$

However, a MIL model learned with this definition has two clear disadvantages: it has a high possibility of increasing the probability of vanishing gradients, and the maximum operator only finds the most discriminative region or instance while lacking the learning of other relevant areas. Another generally used bag label modeling operator is average pooling, which also has a drawback in that it comparably considers all the instances and regions while ignoring the interference of the background noise.

In our work, we proposed a weighted average pooling (WAP) classifier to comprehensively and adaptively integrate the bag representation and calculate the bag probability with a more intensive mechanism. Furthermore, to prevent the relevant boundary information from losing during the down-sample process, we first mirror-padded the input image to transfer its boundary information to the center region and then overlap-croped it to patches of 342×342 pixels. As shown in the left side of Fig. 3.5, we center-croped the image from BreakHis dataset to 684×456 pixels before the mirror padding to make it more regular for the following process and finally produce 6 patches that compose a bag as multiple instances. For Camelyon16 patch-based dataset, we generated a bag including 4 instances. The image-level labels of both datasets were used as the labels for corresponding bags. As a result, we obtained a multiple-instance dataset $\Omega = \{(X_1, Y_1), (X_2, Y_2), \ldots (X_N, Y_N)\}$.

Feature Extractor and Decision Maps

We adopted the convolutional layers of VGG-11 (Simonyan and Zisserman, 2014a) pre-trained by ImageNet dataset as the feature extractor (i.e., FEM) of our network, where all the convolutions were set to valid mode. Fig. 3.5 shows the architectural details. It has five convolutional blocks, the first and second of which are composed of a 3×3 convolution kernel, while each of the other three blocks consists of two 3×3 convolution kernels. Max-pool operation is followed after every convolutional block to halve the feature map resolution for expanding the receptive field. The filter number of each kernel in the five blocks, on the other hand, is doubled with the resolution reduction. A ReLU activation function is added after each convolution operation to introduce nonlinearity to the model and avoid the gradient vanishing problem.

After obtaining feature maps from the fifth convolutional block, we used a 1×1 convolution kernel with two dimensions to integrate the global semantic information and map it to the space corresponding to the class. We named the resulting feature maps decision maps (a.k.a. categories confidence maps). The decision maps have two channels, where the first channel (channel 0) is used to generate the score for the category "Normal" or "benign" and the second one (channel 1) is used to calculate the score for the category "Tumor" or "malignant".

Weighted Average Pooling

In order to comprehensively consider all instances and to increase the network's focus on ROI as well as suppress the background activation, we fused the features of multiple

instances and modeled them into bag-level probability through a WAP (i.e., weighted average pooling) classifier.

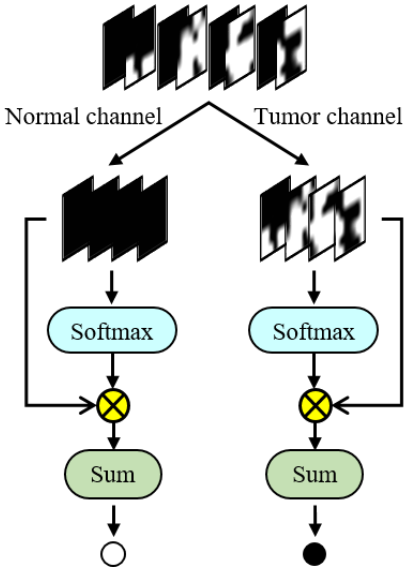

Figure 3.6: The detail of weighted average pooling (WAP) classifier.

Fig. 3.6 provides the detail of our WAP (i.e., weighted average pooling) classifier. The normal and tumor channel features, i.e., f_n and f_t, in decision maps are first passed to a Softmax function separately to calculate the weight coefficient matrix ξ_n and ξ_t of each category feature map over all instances:

$$\xi_i = Softmax\left(f_i\right) \tag{3.3}$$

where $i = n$ is for ξ_n and $i = t$ is for ξ_t. The elements in ξ_n or ξ_t determine the weight of the corresponding pixels in the input image and their summation is equal to 1.

The weight coefficients are then multiplied with their corresponding feature maps through the Hadamard product "$\circ$", and the resulting weighted feature maps are spatially summed over all instances according to different channels to produce the score of each class, i.e., normal or tumor, followed by a Softmax function to calculate the final bag probability $\hat{y}$:

$$\hat{y} = Softmax\left(Sum\left(\xi_n \circ f_n\right), Sum\left(\xi_t \circ f_t\right)\right) \tag{3.4}$$

The gradients for updating convolution parameters in the DMG during the backward pass can be calculated as:

$$\begin{aligned} \frac{\partial\left(F\left(\xi f\right)\right)}{\partial W} &= \frac{\partial\left(F\left(\xi f\right)\right)}{\partial\left(\xi f\right)} \frac{\partial\left(\xi f\right)}{\partial W} \\ &= \frac{\partial\left(F\left(\xi f\right)\right)}{\partial\left(\xi f\right)} \left(\xi \frac{\partial f}{\partial W} + \frac{\partial \xi}{\partial W} f\right) \end{aligned} \tag{3.5}$$

where we use $F\left(\xi f\right)$ to represent Eq. 3.4 for simplicity. The $\xi \frac{\partial f}{\partial W}$ inside the parentheses are scaled with ξ. Since ξ and f are positively correlated (the former is the normalization of the latter), the weight coefficients are adaptive with the update of parameters

in MICNet. That encourages gradients in the DMG originating from target-irrelevant regions to be down-weighted while the gradients from target-relevant regions are up-weighted, which happens to gradients of the FEM as well.

Therefore, the WAP has three main effects on our MICNet. Firstly, it calculates the weight coefficients over all instances, which maps multiple instances to the same plane (or coordinates) and allows the MICNet to learn to avoid its response differences to the same target region in these instances. Secondly, it identifies the most discriminative regions with high weights and keeps the necessary smooth regions with average pooling. Finally, the weight coefficients boost the neuron activations of target regions for both normal and tumor inputs during the forward and backward pass. Especially for tumor input, the weight coefficients not only encourage the activations of lesion tissues but also discredit the activations of healthy tissues.

Explanation Map

Since the output scores of the network are obtained based on the WAP of the decision maps, the calculated weight coefficients are therefore able to reflect the degree of correlation between their corresponding positions in the original image and the target region to some extent. In other words, they can in a way reflect the presence or absence of the target region (i.e., tumor) in the input image, based on which we can use the decision maps to generate explanation maps that explain why the network considers the input image as normal or abnormal.

Considering both the target region and the background, we first up-sample the decision maps to the input instance size and then perform global normalization and binarization to all feature channels to obtain the tumor channel binary maps (see the top right of Fig. 3.5), i.e., the explanation maps, where value 0 means background and value 1 indicates tumor region.

Loss Function

We updated the parameters of our MICNet model by minimizing a binary cross-entropy loss function, which had no access to the pixel-wise annotations:

$$L\left(p, g\right) = -\log \frac{\exp\left(p_g\right)}{\sum_{c=0}^{1} \exp\left(p_c\right)} \tag{3.6}$$

where p and g are the network prediction and ground truth, respectively.

3.3.3 Experimental Settings

Implementation Details

In our experiment, the parameters of our MICNet model were updated by using the Adam optimizer with a batch size of 1. The L2 regularization, i.e., $weight\_decay = 5 \times 10^{-4}$, was adopted to avoid over-fitting. Moreover, the fine-tune strategy (inspired by (Samala et al., 2018)) that freezes the first transferred convolution layer and allows the

rest of the layers to be updated was used. The learning rates for the fine-tuning layers and the last layer were initialized with 10^{-5} and 10^{-3}, respectively. Training sets were augmented by random flip, random color jittering ($brightness$ = 0.5, $contrast$ = 0.5, $saturation$ = 0.5 and hue = 0.05), and random rotation at 90°, 180°, and 270° before being fed into the model.

Evaluation Metrics

The classification performance of our method on the Breakhis test set was evaluated in terms of two commonly used metrics, i.e., patient recognition rate (PRR) and image recognition rate (IRR). The PRR is formulated as:

$$PRR = \frac{1}{N}\sum_{p=1}^{N} PS_p \tag{3.7}$$

where

$$PS_p = \frac{N_{rec}^p}{N_p} \tag{3.8}$$

is the patient score (PS) that defines the ratio of correctly classified images N_{rec}^p to the total number of images N_p for patient p. N is the total number of patients. The IRR is defined as:

$$IRR = \frac{N_{rec}}{N_{img}} \tag{3.9}$$

where N_{rec} represents the number of correctly classified images, and N_{img} is the total number of input images.

For the evaluation of our MICNet in explanation performance on the Camelyon16 test set, we mainly give a visual comparison of the explanation map provided by our MICNet with the pixel-wise ground truth in this chapter.

3.3.4 Results and Discussion

Classification Results on BreakHis Dataset

We compared our MICNet with several state-of-the-art classification methods that were evaluated on the BreakHis Dataset. They are Spanhol et al. (Spanhol et al., 2016a), Bayramoglu et al. (Bayramoglu et al., 2016a), Qi et al. (Qi et al., 2019), A-MIL (Ilse et al., 2018a; Patil et al., 2019), Sudharshan et al. (Sudharshan et al., 2019), and SMSE (Sun et al., 2021), respectively.

The comparison results are given in Table 3.1, where the values marked in bold represent the best performance. It is observed that our MICNet outperforms all the other models in terms of PRR and IRR at all four magnification factors. In particular, the proposed MICNet has a 0.87%, 0.2%, 0.33%, and 0.89% improvement in PRR at these four magnification factors, respectively, compared to the second-best performance. For IRR, our MICNet has a 4.72%, 1.36%, 0.72%, and 2.95% improvement at four magnification factors, respectively.

Table 3.1: Classification results (in %) on BreakHis test set. The best performance is shown in bold. N/A denotes the value that was not provided in the original work.

Methods	40×		100×		200×		400×	
	PRR	IRR	PRR	IRR	PRR	IRR	PRR	IRR
Spanhol et al. (Spanhol et al., 2016a)	90	85.6	88.4	83.5	84.6	82.7	86.1	80.7
Bayramoglu et al. (Bayramoglu et al., 2016a)	83.08	N/A	83.17	N/A	84.63	N/A	82.1	N/A
Qi et al. (Qi et al., 2019)	91.26	89.29	93.1	90.95	92.84	91.61	92.3	90.36
A-MIL (Patil et al., 2019)	N/A	82.95	N/A	86.45	N/A	86.56	N/A	84.43
Sudharshan et al. (Sudharshan et al., 2019)	92.1	87.8	89.1	85.6	87.2	80.8	82.7	82.9
SMSE (Sun et al., 2021)	87.51	N/A	89.12	N/A	90.83	N/A	87.1	N/A
Ours	**92.97**	**94.01**	**93.3**	**92.31**	**93.17**	**92.33**	**93.19**	**93.31**

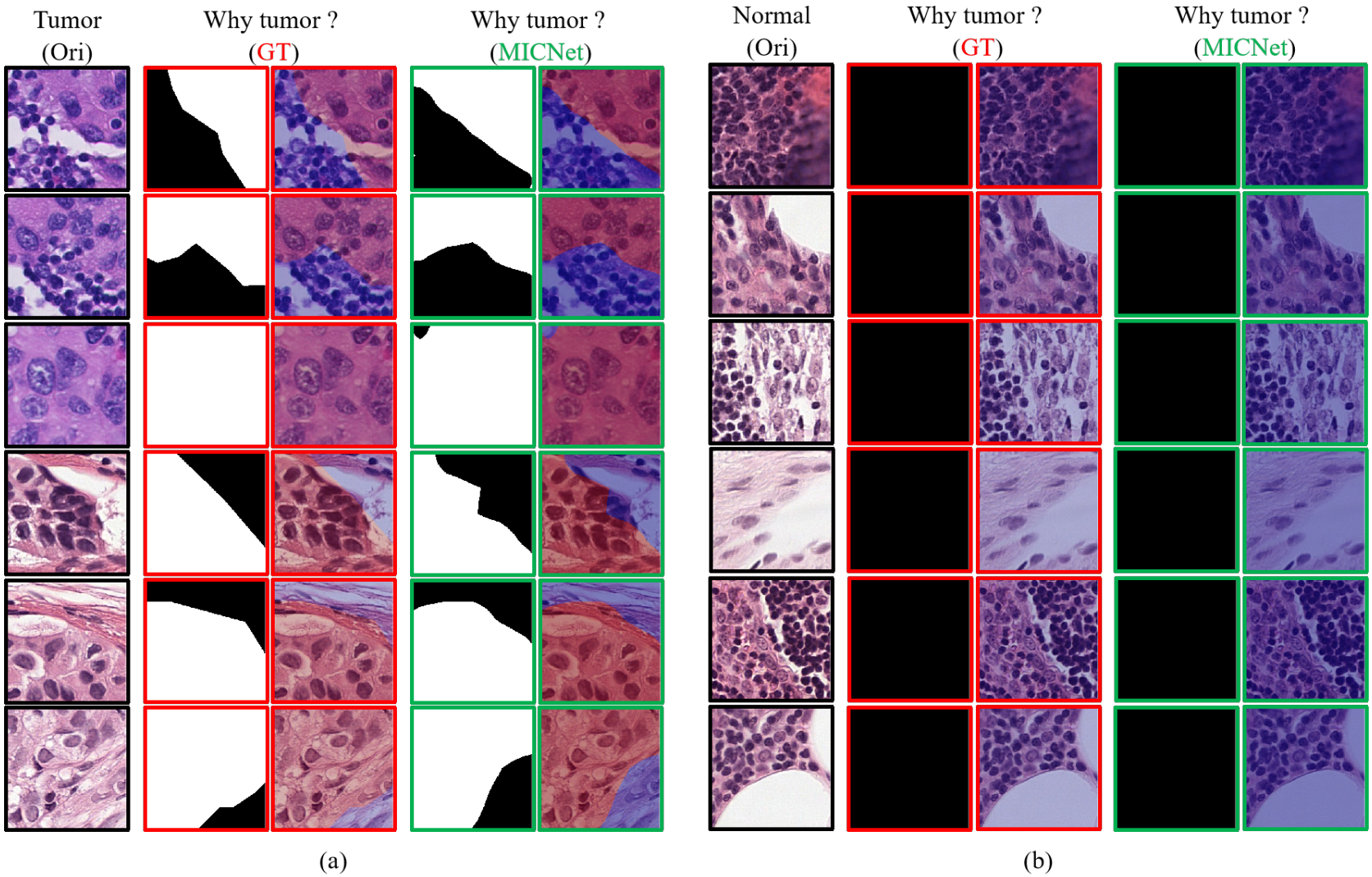

Figure 3.7: Visual explanation of images classification with (a) and without (b) tumors on Camelyon16 patch-based dataset. Red and blue regions in images are tumors and healthy tissues, respectively. Images inside the black, red and green boxes are original images, the ground truth, and explanation map results from our model, respectively.

We believe that better performance was achieved by the well-designed MIL strategy that adopts mirror padding and overlap cropping to avoid information loss. The fine-tuning strategy that maintains the transferred parameters of the first convolution layer prevents the network from over-fitting to the training set. Furthermore, the WAP classifier that models the bag-level probability enhances the network's ability to identify the most discriminative regions relevant to the target while denigrating the background's activations.

Explanation Results on Camelyon16 Patch-Based Dataset

We also performed our method on the Camelyon16 patch-based dataset, which achieved more than 96% IRR on the test set. The visual explanation maps that explain why the network classifies the input as "Tumor" or "Normal" are shown in Fig. 3.7. Fig. 3.7(a) provides the visual explanation of input images with tumors, and Fig. 3.7(b) that of normal images. The first column shows the original images, the second and third columns give the localization of tumors based on the ground truth, and the last two columns provide the explanation maps generated through our method. The binary mask, where value 1 means tumor region and value 0 indicates background, is the computer-level understandable explanation. The heat map, where the tumor region is marked in red and depicted on the right of each binary mask, is the human-level understandable explanation, and it was generated by mapping the binary mask to the original image.

The results show that the explanation maps provided by the proposed MICNet are consistent with the image-level predictions, namely, the explanation maps with tumor regions support the predictions of the tumor class, and the ones without any tumor regions support the normal counterparts. Moreover, the marked tumor regions based on our network's explanation maps are very close to the localization from the ground truth.

3.3.5 Conclusion

In this study, we used the FEM-DMG-classifier configuration (MICNet) of our ConfigNet based on 1×1 convolution kernel (basic DMG) to generate the explanation map for the visually explainable classification of histopathological breast cancer images. We used the convolution layers of the VGG11 model pre-trained by ImageNet as our feature extractor. The MIL, mirror padding and overlap cropping strategies were adopted to avoid information loss and increase classification accuracy. Furthermore, a new WAP was designed to encourage the network to focus more on the target region and generate a good explanation map.

Extensive experiments showed that the proposed MICNet outperforms the state-of-the-art deep learning approaches in terms of both PRR and IRR on the BreakHis dataset. The results on the Camelyon16 patch-based dataset showed that our MICNet is able to accurately classify breast cancer while providing a reliable visual explanation that is of great importance for human understanding and clinical diagnosis in practice.

In the future, further reproducible experiments on other datasets would be helpful for the confirmation of the network's performance, and further comparisons of the

network with other state-of-the-art networks in terms of interpretability might be interesting.

3.4 Breast Tumor Segmentation via the Encoder-Decoder Configuration of Our ConfigNet

3.4.1 Methodology

Fig 3.8 illustrates the overall architecture of the SCAFFNet, of which the encoder is constructed using the first five convolution units (see more details in table 3.2) of VGG16 trained through the ImageNet dataset. The max-pooling indices (i.e., the locations of the maximum feature value in each pooling window) of each down-sampling process of the encoder were saved for the corresponding up-sampling operation in the decoder to alleviate the loss of boundary information relevant to targets. In order to reduce learnable parameters that account for computational complexity while increasing the nonlinearity and network depth that enhances the feature representation, we developed a decoder with "bottleneck" structures. Furthermore, a spatial and channel attention-guided feature fusion module (SCAFFM) was proposed for the skip connection between the encoder and decoder to encourage the network to identify target regions more accurately.

Table 3.2: Encoder structure.

Convolution unit	Structure
1	$[\text{Conv } 3\times3 + \text{BN} + \text{ReLU}, C = 64]\times 2$ $\text{Maxpool } 2\times2$
2	$[\text{Conv } 3\times3 + \text{BN} + \text{ReLU}, C = 128]\times 2$ $\text{Maxpool } 2\times2$
3	$[\text{Conv } 3\times3 + \text{BN} + \text{ReLU}, C = 256]\times 3$ $\text{Maxpool } 2\times2$
4	$[\text{Conv } 3\times3 + \text{BN} + \text{ReLU}, C = 512]\times 3$ $\text{Maxpool } 2\times2$
5	$[\text{Conv } 3\times3 + \text{BN} + \text{ReLU}, C = 512]\times 3$ $\text{Maxpool } 2\times2$

Decoder with Bottleneck Structure

The decoder is shown in the right hand of Fig. 3.8, where we can see that it is almost symmetrical to the encoder except for the extra 1×1 convolution to produce the "bottleneck" structure. First of all, the features of the last layer (including down-sampling) of the encoder are up-sampled through inverse max-pooling using the memorized corresponding max-pooling indices, which generates sparse feature maps with the same resolution that of the maps of the penultimate layer in the encoder. Afterward,

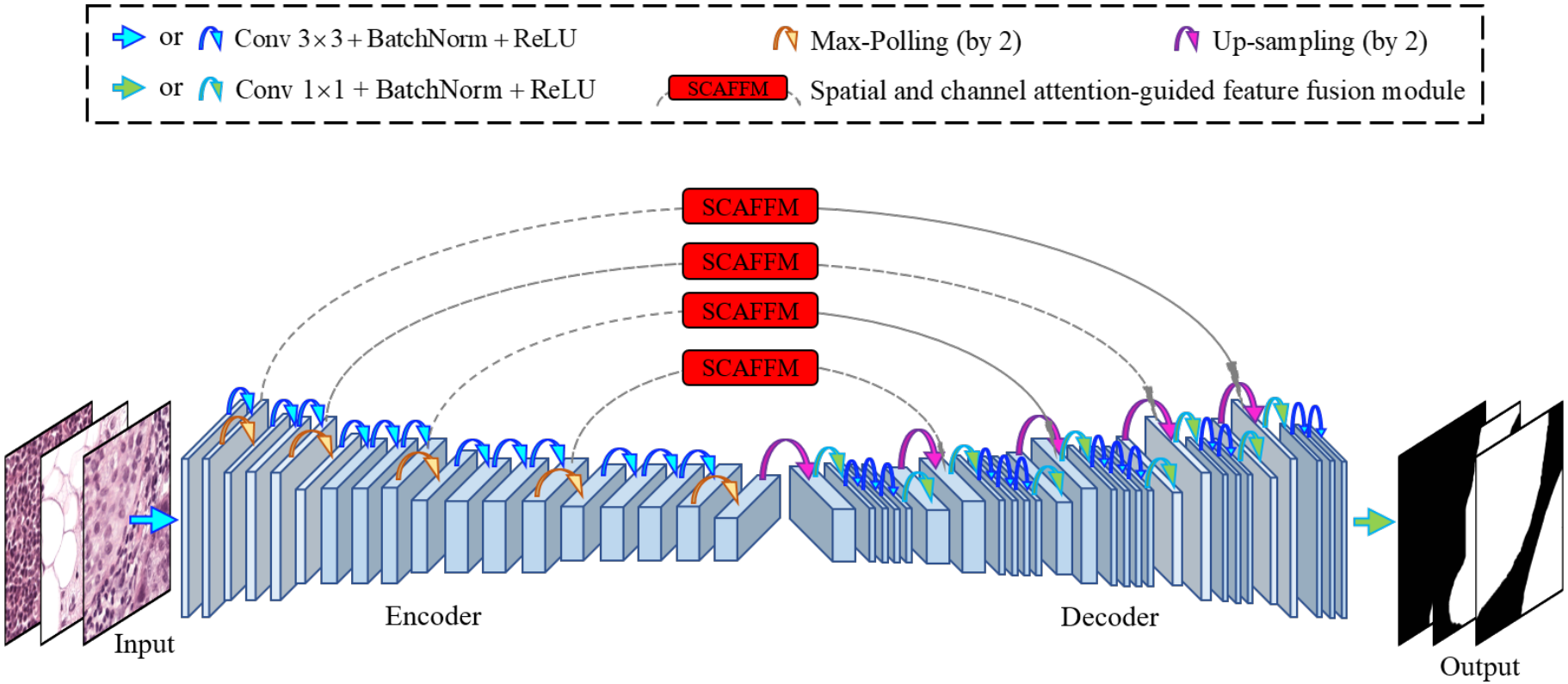

Figure 3.8: Illustration of SCAFFNet.

a 1×1 convolution filter, followed by batch normalization and ReLU non-linear activation function, is adopted to integrate the features in different channels and map them to feature maps with a lower dimension (i.e., fewer channels). The resulting lower-dimensional features are then fed into several trainable convolutional layers (with 3×3 convolution filters) corresponding to the last convolution unit of the encoder, which transforms the sparse up-sampled feature maps into dense maps. To obtain the features having the same dimension as the saved max-pooling indices that will be used in the following up-sampling process corresponding to the encoder, another 1×1 convolution filter is used to increase the channels of the dense feature maps.

The above processes compose a convolution unit with a "bottleneck" structure, and the decoder has a total of five convolution units with the 1×1 convolution filter of the last convolution unit used to generate the segmentation result. Table 3.3 provides more details about these five convolution units, where "Maxunpool" denotes the inverse max-pooling operation. Notably, the inputs following the up-sampling operation of each convolution unit after the first convolution unit of the decoder are generated via the skip connection (achieved by the SCAFFM) between features output from the previous convolution unit and features output from the corresponding convolution unit (before the max-pooling) of the encoder.

Table 3.3: Decoder structure.

Convolution unit	Structure
1	Maxunpool 2×2 Conv 1×1 + BN + ReLU, C = 64 [Conv 3×3 + BN + ReLU, C = 64] $\times 3$ Conv 1×1 + BN + ReLU, C = 512
2	Maxunpool 2×2 Conv 1×1 + BN + ReLU, C = 64 [Conv 3×3 + BN + ReLU, C = 64] $\times 3$ Conv 1×1 + BN + ReLU, C = 256
3	Maxunpool 2×2 Conv 1×1 + BN + ReLU, C = 64 [Conv 3×3 + BN + ReLU, C = 64] $\times 3$ Conv 1×1 + BN + ReLU, C = 128
4	Maxunpool 2×2 Conv 1×1 + BN + ReLU, C = 32 [Conv 3×3 + BN + ReLU, C = 32] $\times 2$ Conv 1×1 + BN + ReLU, C = 64
5	Maxunpool 2×2 Conv 1×1 + BN + ReLU, C = 32 [Conv 3×3 + BN + ReLU, C = 32] $\times 2$ Conv 1×1, C = 1

Spatial and Channel Attention-Guided Feature Fusion Module

The SCAFFM (i.e., spatial and channel attention-guided feature fusion module) structure is provided in Fig. 3.9, where "same" refers to the convolution operation keeping

the input resolution. It is mainly composed of two attention mechanism branches, i.e., spatial attention branch and channel attention branch, followed by an attention-guided feature fusion layer.

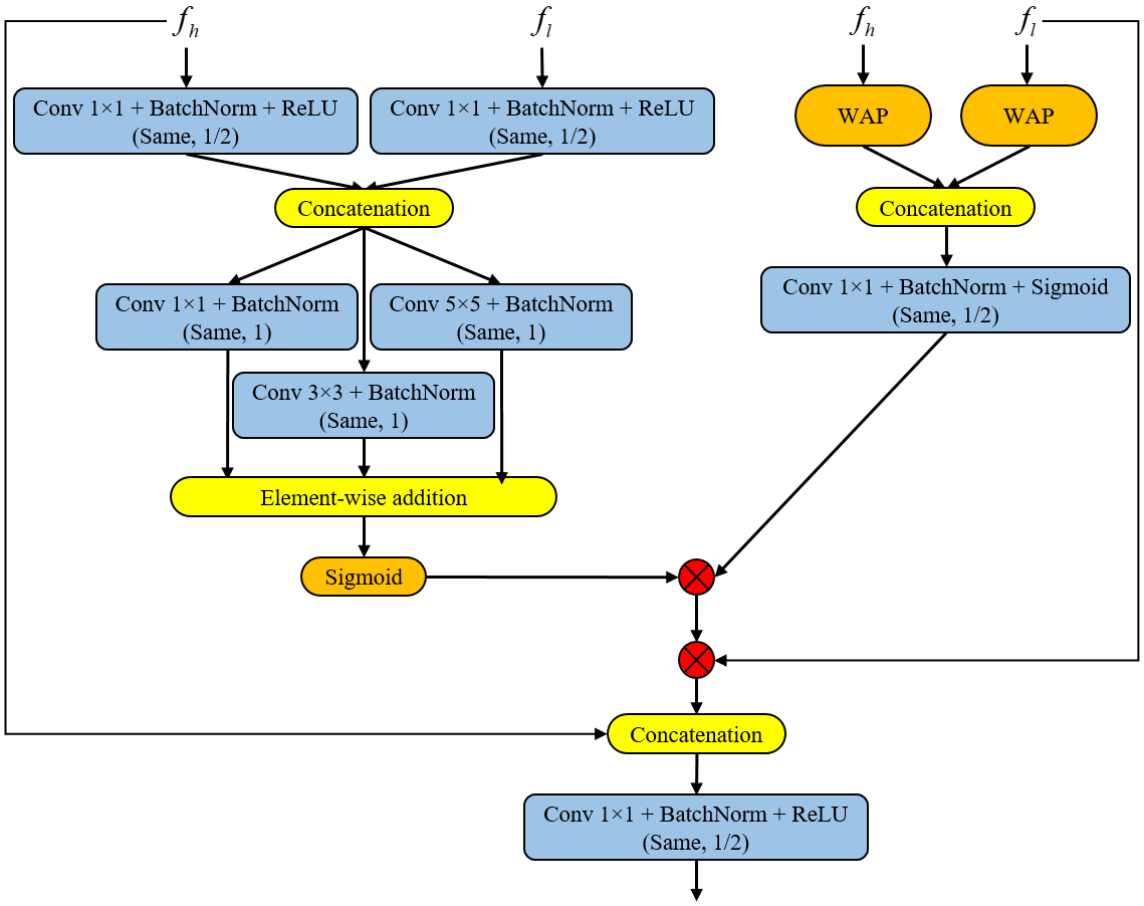

Figure 3.9: The detail of SCAFFM architecture.

For the spatial attention branch, the high-level (or deep) features f_h from the decoder as well as the low-level (or shallow) features f_l from the encoder are first fed into a 1×1 convolution filter, respectively, to halve the number of feature channels, which reduces the requirement of convolution parameters after the following concatenation of the high-level and low-level features. Batch normalization and ReLU activation function are added closely after each convolution filter. The concatenated features are then passed into a multi-size convolutional layer that is constructed with 1×1, 3×3, and 5×5 convolution filters in parallel, followed by an element-wise addition fusion operation. The parallel convolution filters of different sizes are used to increase the diversity of receptive fields so as to enhance the semantic information (e.g., location and boundary) relevant to the target in the spatial attention coefficient map, which is generated by the Sigmoid function after the element-wise addition fusion operation. Each convolution filter in the multi-size convolutional layer is followed by batch normalization.

For the channel attention branch, the high-level and low-level features are first fed into a WAP (i.e., weighted average pooling, see more details in Section 3.3), respectively, to obtain a coarse weight score of the feature map in each channel, followed by the

channel-wise concatenation. A 1×1 convolutional layer (followed by batch normalization) that halves the number of channels is used to generate fine weight scores, which are fed into the Sigmoid function to produce the channel attention coefficient vector.

For the subsequent attention-guided feature fusion layer, the spatial attention coefficient map and the channel attention coefficient vector are first fused to calculate the final attention coefficients, which are then used to filter the low-level features. The fusion and filtering operations of this process are achieved through Hadamard products. Finally, the filtered low-level features are channel-wise concatenated with the high-level features and are then passed into a 1×1 convolutional layer (followed by the batch normalization and ReLU activation function) that halves the number of the channels.

3.4.2 Experimental Settings

Evaluation Metrics

To evaluate the performance of our SCAFFNt, many commonly used evaluation metrics, including mIoU (mean intersection over union), Dice index ($Dice^+$ for foreground and $Dice^-$ for background), Precision, Recall (a.k.a. Sensitivity or true positive rate), Specificity (a.k.a. true negative rate), false positive rate (FPR), and false negative rate (FNR), were used in our experiments. They are formulated as follows:

$$mIoU = \frac{1}{N} \sum_{i=0}^{N-1} \frac{p_{ii}}{\sum_{j=0}^{N-1} p_{ij} + \sum_{j=0}^{N-1} (p_{ij} - p_{ii})} \tag{3.10}$$

$$Dice = \frac{2\,|\mathrm{P} \cap \mathrm{G}|}{|\mathrm{P}| + |\mathrm{G}|} \tag{3.11}$$

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{3.12}$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3.13}$$

$$Specificity = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{3.14}$$

$$FPR = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{3.15}$$

$$FNR = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}} \tag{3.16}$$

where $N = 2$ is the number of pixel categories, i.e., 0 and 1. p_{ii} is the number of correctly predicted pixels. p_{ij} is the number of pixels whose true category is i and predicted category is j. P and G are the binary target (foreground for $Dice^+$ and background for $Dice^-$) masks of the prediction and ground truth, respectively, and $|\cdot|$ the cardinality of a set. The confusion matrix (i.e., TP, TN, FP, and FN) is calculated over all pixels on the test set.

Implementation Details

All the methods evaluated on the Camelyon16 patch-based dataset, including ours, were trained using the Adam optimizer with $\beta 1 = 0.9$, $\beta 2 = 0.999$, a weight decay of 5×10^{-4}, and a batch size of 32 for 224×224 images resized from the 512×512 patches. The learning rates for all the fine-tune layers and other layers without transferred parameters were initialized with 10^{-5} and 10^{-3}, respectively. All the methods were trained using pixel-level ground truth. We used the pixel-wise binary cross-entropy loss to update the parameters in our model and followed the settings in the original works for other methods, including most of their hyper-parameters. We trained all the methods for 90 epochs to find the best convergence. Training set was augmented by random horizontal and vertical flips with a probability of 0.5, random color jittering ($brightness = 0.5$, $contrast = 0.5$, $saturation = 0.5$ and $hue = 0.05$), and random rotation with an angle in $\{0°, 90°, 180°, 270°\}$ before being fed into all the networks.

3.4.3 Results and Discussion

We have compared our SCAFFNet with many state-of-the-art deep learning methods to demonstrate the effectiveness of the encoder-decoder configuration of our ConfigNet and its superior performance in breast tumor segmentation based on the developed SCAFFM. they are U-Net (Ronneberger et al., 2015), SegNet (Badrinarayanan et al., 2017), R2U-Net (Alom et al., 2018), Attention U-Net (Schlemper et al., 2019), and R2AU-Net (Zuo et al., 2021).

Table 3.4 provides the quantitative comparison results of these methods. It is seen that our SCAFFNet outperforms all the other models in terms of mIoU, Dice index of foreground and background, Precision, and Specificity (or FPR). More specifically, SCAFFNet has an approximately 0.79% improvement in mIoU and an approximately 1.54% improvement in $Dice^+$ compared to the second-best method SegNet. The R2U-Net and R2AU-Net have better performance in terms of Recall/Sensitivity (or FNR), with almost 6.14% and 4.48% improvement (or drop), respectively, compared to our SCAFFNet, but their mIoU, $Dice^-$, Precision, and Specificity (or FPR) are all more than 8.5% lower (or higher). Particularly, our SCAFFNet has an approximately 16.27% and 14.11% improvement (drop) in Specificity (or FPR) compared to the R2U-Net and R2AU-Net.

Furthermore, as observed in Fig. 3.10, the P-R curve of our SCAFFNet tends to be on the top right and has a higher Precision value over a wide threshold range. The ROC curve of the SCAFFNet provided in Fig. 3.11 tends to be on the top left and has a higher true positive rate (TPR) over a wide threshold range. These further verify the superior performance of our SCAFFNet over other segmentation deep learning models. The visual results of the comparison are shown in Fig. 3.12. It shows that our SCAFFNet has segmentation results much closer to the ground truth. Other methods either present more severe over-segmentation (i.e., more false positives), such as samples in the fifth row, or more dramatic under-segmentation (i.e., more false negatives), such as samples in the last row. In particular, It is observed that our SCAFFNet generates more elaborate segmentation results for tumor tissues with more complex boundaries, such as samples in the first, third, and fourth rows.

Table 3.4: Comparison results (in %) of segmentation performance of different methods on Camelyon16 patch-based test set. The best performance is indicated in bold. Sen. and Spe. are simplified from Sensitivity and Specificity, respectively.

Methods	mIoU	$Dice^+$	$Dice^-$	Precision	Recall/Sen.	Spe.	FPR	FNR
U-Net (Ronneberger et al., 2015)	76.03±0.44	70.98±0.72	88.58±0.35	76.91±1.40	72.11±0.80	91.12±0.56	8.88±0.56	27.89±0.80
SegNet (Badrinarayanan et al., 2017)	77.22±0.31	72.40±0.89	89.23±0.13	78.63±0.74	72.92±1.20	92.08±0.35	7.92±0.35	27.08±1.20
R2U-Net (Alom et al., 2018)	67.35±2.97	71.48±1.08	79.30±3.26	69.57±3.22	**80.61±2.86**	76.23±4.76	23.77±4.76	**19.39±2.86**
Attention U-Net (Schlemper et al., 2019)	76.31±0.34	71.55±0.77	88.79±0.30	77.85±1.57	72.06±0.87	91.48±0.69	8.52±0.69	27.94±0.87
R2AU-Net (Zuo et al., 2021)	68.39±2.34	70.68±0.55	80.53±2.84	70.31±3.55	78.95±3.54	78.39±4.24	21.61±4.24	21.05±3.54
SCAFFNet	**78.01±0.32**	**73.94±0.65**	**89.54±0.27**	**78.84±0.84**	74.47±1.20	**92.50±0.51**	**7.50±0.51**	25.53±1.20

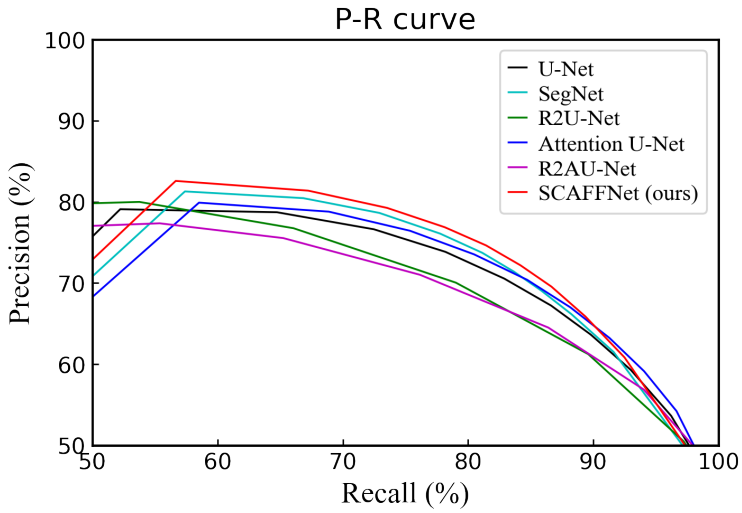

Figure 3.10: P-R curve of the comparison.

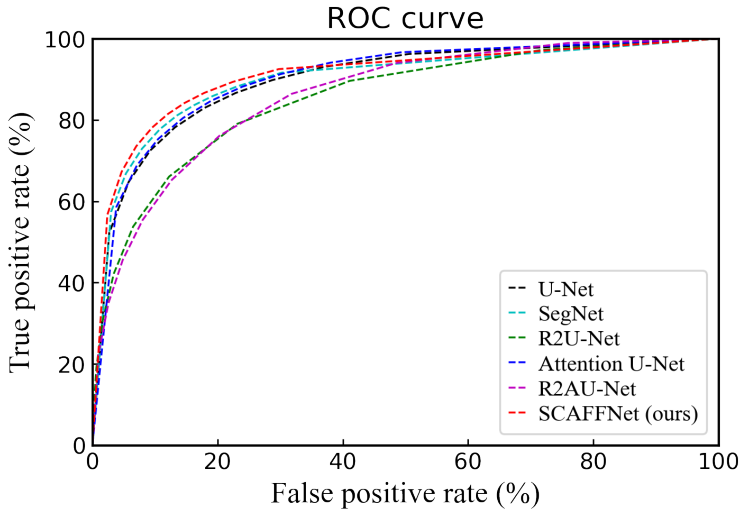

Figure 3.11: ROC curve of the comparison.

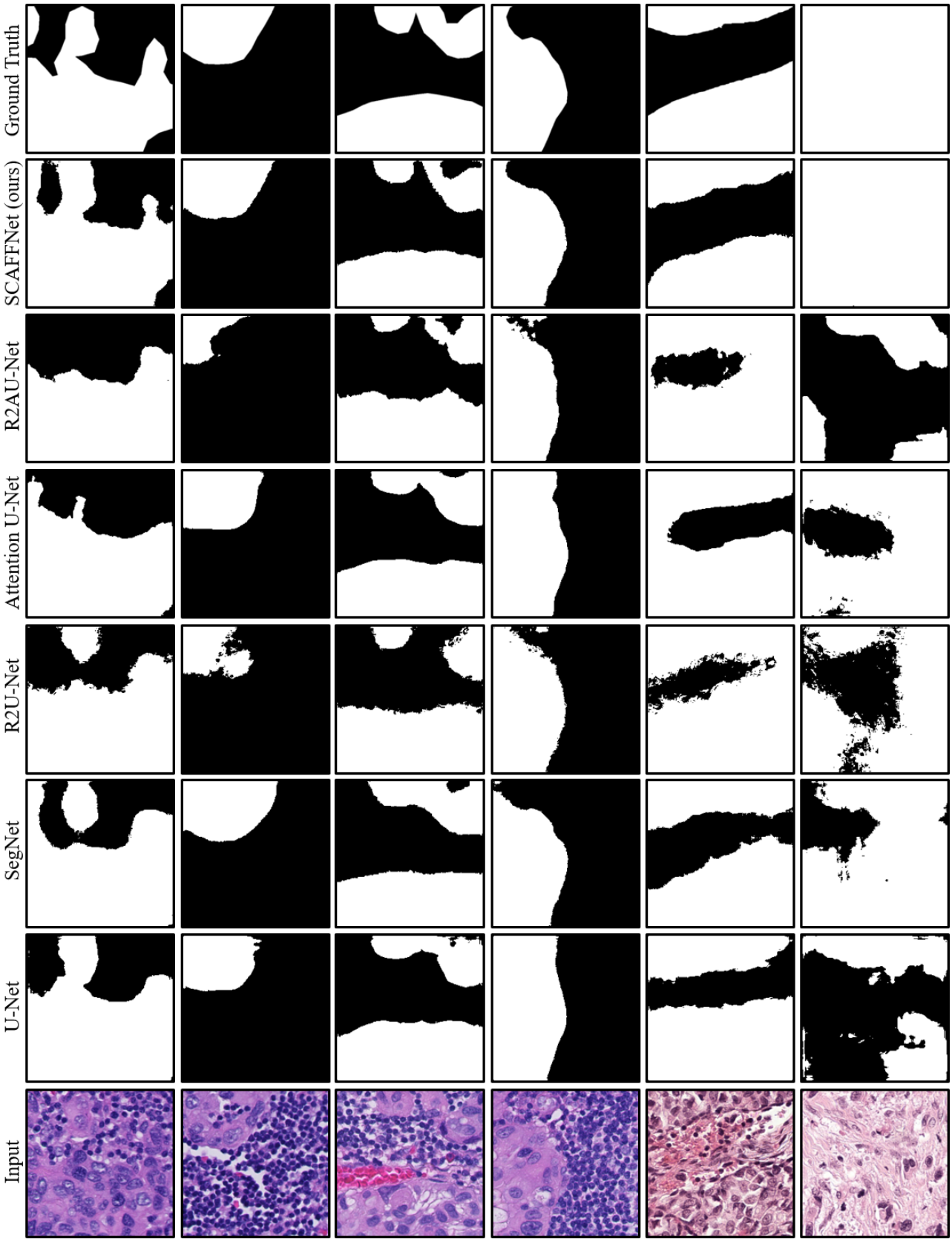

Figure 3.12: Some visual results of the comparison.

3.4.4 Conclusion

In this study, we used the encoder-decoder configuration of our ConfigNet based on SCAFFM and a "bottleneck" decoder for the tumor segmentation of histopathological breast cancer images. The convolution layers of the VGG16 pre-trained by ImageNet were adopted as our encoder. The SCAFFM, with a spatial and channel attention mechanism that guides the skip connection of features between the encoder and decoder, was developed to encourage the network to separate foreground and background regions more accurately. A decoder with the "bottleneck" structure developed to increase the network depth, thus improving the feature representation, was used to generate segmentation results.

Experimental results showed that the proposed SCAFFNet outperformed the state-of-the-art deep learning models in terms of many segmentation evaluation metrics (i.e., mIoU, Dice index of both foreground and background, Precision, and FPR) on the Camelyon16 patch-based dataset. The visual results showed that our SCAFFNet was able to segment breast tumors with complex boundaries more accurately than the state-of-the-art methods, suggesting its promising potential in clinical usage.

3.5 Summary

This chapter mainly describes the key methodology developed for this thesis:

- First, we presented a detailed description of the proposed configurable convolutional neural network (ConfigNet), including the overall architecture of the ConfigNet, the detailed structure (including the supervision strategy) of the FEM-DMG-classifier configuration devoted to explainable classification of medical images, and the detailed structure of the encoder-decoder configuration dedicated to object segmentation or localization.

- Secondly, we conducted extensive experiments for the FEM-DMG-classifier configuration (named MICNet) based on VGG11, basic DMG, and WAP classifier on two publicly available breast cancer datasets and compared it with many existing deep learning methods to evaluate the classification performance of our ConfigNet on histopathological breast cancer images. The quantitative comparison results on the BreakHis dataset demonstrated the superior classification performance of the MICNet, and the visual results of the MICNet on the Camelyon16 patch-based dataset verified its effectiveness in providing a reasonable explanation of the classification.

- Finally, we implemented a comparative study of the encoder-decoder configuration (named SCAFFNet) based on the developed SCAFFM and "bottleneck" decoder with existing state-of-the-art segmentation models to evaluate the segmentation performance of the ConfigNet. The experimental results demonstrated that our SCAFFNet outperforms other segmentation models in terms of several evaluation metrics, including mIoU, Dice index of foreground and background, Precision, and FPR. In addition, the SCAFFNet has a finer segmentation on breast tumors with complex boundaries.

Chapter 4

ExplaCNet: Explainable Classification of Histopathological Breast Cancer Images Based on Weakly-Supervised Learning

4.1 Introduction 104
4.2 Methodology 107
4.2.1 Input Preprocessing 107
4.2.2 Decision Map Generator 107
4.3 Experimental Setting 110
4.3.1 Evaluation Metrics 110
4.3.2 Implementation Details 112
4.4 Results and Discussion 113
4.4.1 Comparison in Performance of Explanation 113
4.4.2 Comparison in Performance of Classification 119
4.4.3 Ablation Study 119
4.5 Conclusions 123

4.1 Introduction

Breast cancer is common cancer that remains the leading cause of death in women worldwide (Sung et al., 2021). Early diagnosis plays a pivotal role in breast cancer treatment and the improvement of the survival rate in female patients. With the development of imaging technology, breast cancer diagnosis now includes digital mammography, magnetic resonance imaging (MRI), ultrasound, biopsy, etc. Biopsy is the gold standard for the diagnosis of almost all cancers (Strayer, 2015), and it is traditionally done by experienced pathologists needing years of training. However, manual diagnosis requires significant time investment for careful inspection and is subject to low inter-rater agreement between experts as well as low intra-rater consistency across multiple readings of the same expert. The insufficiency of experienced pathologists is also a big issue for some hospitals. These drawbacks drove the emergence of cost-effective computer-aided diagnosis (CAD) systems (AlZubaidi et al., 2017; Tey et al., 2018; Yanase and Triantaphyllou, 2019; Chan et al., 2020; Hsu et al., 2021; Dika et al., 2022) that avoid human factors such as fatigue, attention span, and differences in experiences and have the potential to effectively alleviate most of the workload of pathologists, thereby solving these problems.

In recent years, deep learning methods, in particular convolutional neural networks (CNNs) (He et al., 2016b; Chen et al., 2018; Badrinarayanan et al., 2017; Zhang et al., 2018b; Shaham et al., 2019; Ashual and Wolf, 2019; Teed and Deng, 2020; Zhang et al., 2020; Radford et al., 2021; Dhariwal and Nichol, 2021), have made substantial strides and achieved state-of-the-art performances in computer vision and image processing due to their superior properties in feature learning and identification. Inspired by these great successes, researchers started to leverage CNNs to develop CAD systems with high accuracy to achieve the diagnosis of histopathological breast cancer images. For instance, there have been many research efforts on breast tissue classification (Spanhol et al., 2016a; Song et al., 2017; Sudharshan et al., 2019; Kumar et al., 2020; Gour et al., 2020; Sun et al., 2021; Schirris et al., 2022) and lesion tissue segmentation (or localization) (Aswathy and Jagannath, 2017; Li et al., 2018; Krithiga and Geetha, 2020; Jin et al., 2020; van Rijthoven et al., 2021; Schmitz et al., 2021).

Despite the impressive advances and comparable or even superior diagnosis performance (compared to pathologists) that CNNs brought to CAD systems, they still face a substantial challenge for their deployment in real-world clinical settings. CNN models are often discouraged in clinical diagnosis due to their "black box" nature deriving from lacking explainability (also known as interpretability) or transparency in decisions. This nature makes it difficult for end-users to interrogate and understand the diagnosis result, which inevitably causes user skepticism and distrust. Human experts need certain visual clues, such as cell morphology, texture, and structure, to make correct decisions for histopathological images. Therefore, a CAD system (i.e., CNN model) that can provide a reliable visual explanation exhibiting explicit domain-specific evidence to support its decision is fundamental for its deployment in actual clinical settings, where the transparency (i.e., providing visual evidence for the decision) is a potential liability for deep learning methods.

In fact, many promising efforts having the potential to introduce the mechanisms of transparency, understandability, and explainability into CNN models have already

been made recently; methods based on weakly-supervised segmentation and localization that are only trained with image-level annotations are the most widely studied hot spots. These methods mainly produce the visual explanation map (i.e., explainability) based on the discriminative regions (or up-weighted gradients) that contribute to the final decision. They are generally divided into two categories, i.e., model-specific and model-independent (also known as model-agnostic). Model-specific methods produce the visual explanation through the forward inference of an individual model, such as CAM (Zhou et al., 2016), WILDCAT (Durand et al., 2017), ACoL Zhang et al. (2018a), SCOUTER (Li et al., 2021b), D-RISE (Petsiuk et al., 2021), multiple instance learning (MIL)-based methods (Ilse et al., 2018b; Patil et al., 2019), and other methods based on weakly-supervised learning (WSL) (Singh et al., 2020; Ciga and Martel, 2021; Belharbi et al., 2021; O'Shea et al., 2022). Model-independent methods, however, generate their visual explanation via the backward inference from results to causes and is not restricted to specific network architecture. It is primarily applicable to post-hoc analysis, such as guided back-propagation (Springenberg et al., 2014) and methods built on CAM, e.g., Grad-CAM (Selvaraju et al., 2017a), Grad-CAM++ (Chattopadhay et al., 2018), Ablation-CAM (desai and Ramaswamy, 2020), and Score-CAM (Wang et al., 2020).

Particularly, many explainable CNN models have been investigated for the explainable classification of breast cancer images to facilitate the deployment of deep learning-based CAD systems in real-world clinical settings. Model-independent methods are more frequently applied and investigated due to their generality. Huang and Chung (2019) proposed a CELNet model with multi-branch attention modules and a deep supervision mechanism to address the difficulty of classifying histopathological breast cancer images. After obtaining the classification result, they used the combination (CELM) of Grad-CAM (Selvaraju et al., 2017a) and guided back-propagation (Springenberg et al., 2014) to generate the evidence localization to explain the decision. Another study (Adoui et al., 2020) also employed the Grad-CAM to provide a visual explanation to support the decision of their network that predicts breast tumor response to chemotherapy using quantitative MR images. Kaplun et al. (2021) adopted Zernike image moments (Khotanzad and Hong, 1990) to extract complex features from breast cancer histopathological images, and they used simple neural networks to generate classification results whose explainability was generated through the local interpretable model-agnostic explanations (LIME) method (Ribeiro et al., 2016). Rodriguez-Sampaio et al. (2022) first used EfficientNet (Tan and Le, 2019) to achieve the classification of mammographic breast cancer images and then studied the classification interpretability of Occlusion Sensitivity (Zeiler and Fergus, 2014), SmoothGrad (Smilkov et al., 2017), Integrated Gradients (Sundararajan et al., 2017), GradCAM, GradCAM++ (Chattopadhay et al., 2018), and ScoreCAM (Wang et al., 2020) on the test set. For model-specific methods, Joshi et al. (2020) designed an encoder and decoder architecture to address the problem of explainable classification of mammographic breast cancer images. The encoder was adopted to produce a classification output while the decoder was used to produce the rough segmentation maps, which were combined with image-level labels to encourage the CAM (Zhou et al., 2016) trained with a novel methodology to generate a better explanation, i.e., heat map. Shen et al. (2021) proposed a framework named GMIC to classify high-resolution breast cancer

screening mammograms. The authors used a global module with low capacity yet memory-efficient to extract the global context of the whole image and obtain saliency maps providing coarse localization of possible target regions. They then used a local module with a higher capacity to generate a refined visual explanation from the coarse saliency maps and a fusion module to generate the final classification prediction.

However, there are great challenges ahead of the considerable progress made in explainable classification. Because of having no access to the network forward inference, the model-independent method mainly explains why the network makes a certain decision without regard to its correctness. It also hardly satisfies the constraint that target regions indicated by a clinically human-understandable visual explanation map should highlight the lesion tissue as comprehensively as possible and have fewer false positives. Satisfying this constraint is the precondition of a clinically reliable and logical explanation map. The model-specific method struggles with a crucial conflict between better classification performance and a better explanation map, although it has a closer connection to the network's decision. This conflict is particularly prominent in the clinical classification of histopathological images with less inter-class differences between tumor and normal tissues and non-ignorable intra-class differences between tumors.

Considering the necessity of meeting the visual explanation constraint in real-world clinical settings and the great importance of the logical connection between explanation and decision, we proposed an explainability-favored FEM-DMG-classifier configuration of our ConfigNet. The FEM-DMG-classifier configuration, named ExplaCNet, is a model-specific explainable method. It can solve the aforementioned issue and achieve the visually explainable classification of histopathological breast cancer images. We demonstrated the exceeding performance of our method in the clinically relevant diagnosis of histopathological breast cancer images through comprehensive experiments on the public Camelyon16 patch-based dataset and BreakHis dataset. Furthermore, we reported the ablation study to evaluate and verify the effectiveness of the main components used in our ExplaCNet. Our main contributions are as follows:

- We proposed a new explainable framework, i.e., ExplaCNet, based on WSL to mimic the diagnosis logic of human experts by providing visual explanation maps that indicate the localization and regions of lesion tissues. It follows the transparency criteria closely relevant to the safety, ethics, and reliability in clinical diagnosis and provides considerable support for the deployment of deep-learning model-based CAD systems in real-world clinical settings.

- We adopted the multiple instance learning (MIL) strategy to enhance our network's ability to deal with images of big sizes and developed an adaptive weighted average pooling (WAP) classifier to fuse multiple instances and model the final bag probability. The MIL strategy not only improves the performance of our ExplaCNet in normal tissue identification and image classification but also opens up the possibility of the network for processing whole-slide pathological images with billions of pixels. The adaptive WAP classifier allows the network to recognize more lesion regions, which decreases false negatives.

- We developed a DMG composed of multi-scale filters to encourage the network

to make a reasonable compromise between the sensitivities to tumor and normal tissues, thus improving the network's performance in identifying tissues of both categories and finally generating a decision map with lesion and normal regions separated more correctly for classification and explanation.

4.2 Methodology

The architecture of our ExplaCNet including input image preprocessing is provided in Fig. 4.1. The proposed method has three key components, i.e., MIL, DMG, and WAP.

During the diagnosis procedure as well as the training, the input images were first cropped into a bag of instances which were then input into a FEM to obtain feature maps containing enough contextual and discriminative information relevant to the target. Afterwards, a DMG was used to deal with the encoded deep features and generate a decision map that indicates the presence of targets and their localization. Finally, the resulting decision map was fed into a WAP classifier followed by a Softmax function to produce the bag probability, i.e., the classification probability of the input image.

Note that the FEM used in our method was built on the VGG16 architecture (without batch normalization due to our small batch size, i.e., 1) (Simonyan and Zisserman, 2014b) pre-trained with ImageNet dataset; we used the first fifth layer (out of a total of six layers including the classifier layer) of this model to encode input images. We have described MIL and WAP classifier in detail in Chapter 3 and we will give a detailed description of the proposed DMG in this chapter.

4.2.1 Input Preprocessing

In order to produce an input with a bag of instances, we used a 256×256 window to crop the input image into several patches (varies between datasets), where the image label is considered the bag-level label, as shown in the left side of Fig. 4.1. "Or" means different datasets were trained and tested separately.

4.2.2 Decision Map Generator

For a classification network, the feature maps output from the FEM must contain the most discriminative information to allow the following classification layer to recognize different categories, where some smooth features of the target are discarded. It is why the max-pooling operator is prevalently used in classification networks. In our task, however, correct classification alone is not enough since we need to provide a human-understandable explanation map that contains as many target regions as possible to explain the decision. It would be problematic if we only used max-pooling in our network, as these discarded smooth features can be used to encourage the network to extract more target regions. Here, to better satisfy the constraint that target regions indicated by a clinically human-understandable visual explanation map should highlight the lesion tissue as comprehensively as possible and have fewer false

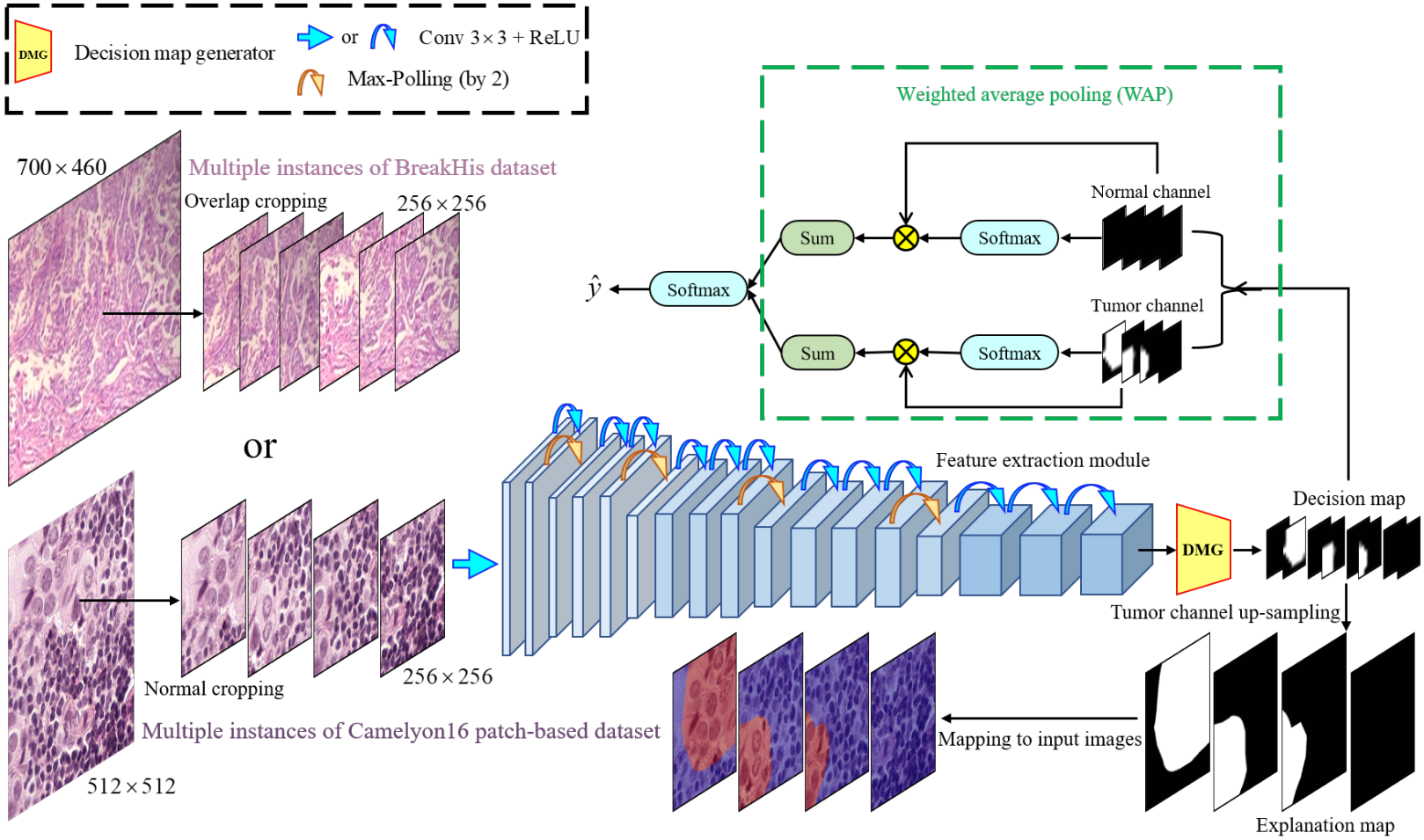

Figure 4.1: An illustration of ExplaCNet architecture.

positives, we proposed a DMG (i.e., decision map generator) module on top of the MICNet configuration from the ConfigNet to address this issue.

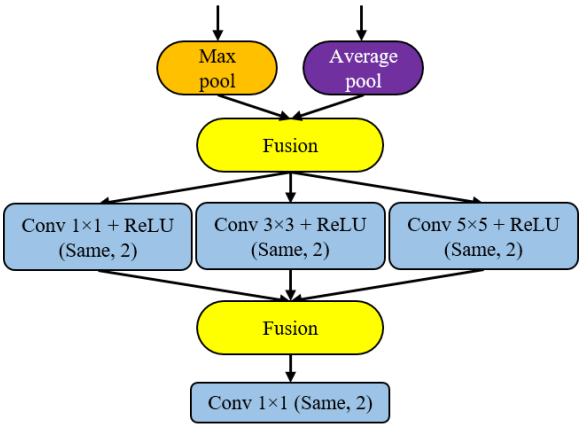

Figure 4.2: The architecture of decision map generator (DMG) module. “Same" means that the convolution keeps the resolution. “2" denotes that the output has two channels.

Fig. 4.2 shows the detailed architecture of the DMG. At first, we used a parallel max-pooling and average pooling layer to down-sample the feature maps f_{FEM} from the FEM, which identifies both the most discriminative features and the smooth features of the target. These features were then fused via concatenation:

$$f'_{\text{FEM}} = \left(cat\left(MaxPool\left(f_{\text{FEM}}\right), AveragePool\left(f_{\text{FEM}}\right)\right)\right) \tag{4.1}$$

where f'_{FEM} is the fused features and cat denotes the channel-wise concatenation.

Afterwards, f'_{FEM} was passed to a multi-scale convolutional layer for decryption to generate coarse categories confidence maps corresponding to the categories. As shown in Fig. 4.2, the multi-scale convolutional layer is composed of 1×1, 3×3, and 5×5 convolutions in parallel, which was inspired by Inception architecture (Szegedy et al., 2015b). The resulting features are therefore calculated from different receptive fields that cover diverse contextual information. These features were later fused through concatenation as well:

$$f_{\text{Multi}} = \left(cat\left(\Phi_{1\times1}(f'_{\text{FEM}}), \Phi_{3\times3}(f'_{\text{FEM}}), \Phi_{5\times5}(f'_{\text{FEM}})\right)\right) \tag{4.2}$$

where f_{Multi} denotes the fused features output from the multi-scale convolutional layer, and

$$\Phi_{n\times n}\left(f'_{\text{FEM}}\right) = \text{ReLU}\left(W_{n\times n} * f'_{\text{FEM}} + b\right) \tag{4.3}$$

is the convolution operation with convolution kernel $W_{n\times n}$ and bias b, $n = 1, 3, 5$. $*$ is the convolution operator and ReLU is the activation function.

Finally, f_{Multi} was decoupled and refined to more precise decision maps $\text{Map}_{\text{decision}}$ (i.e., categories confidence maps) indicating target regions through a 1×1 convolutional layer:

$$\text{Map}_{\text{decision}} = W_{1\times1} * f_{\text{Multi}} + b \tag{4.4}$$

where $W_{1\times1}$ denotes the 1×1 convolution kernel. The explanation map, i.e., binary mask, which indicates the presence or absence of lesion tissues, was generated through interpolation of the tumor-channel decision map.

4.3 Experimental Setting

4.3.1 Evaluation Metrics

We adopted many commonly used metrics to quantitatively evaluate our model on visually explainable classification of breast cancer and compare it with different methods. Furthermore, we defined a new metric named “Explainability" to more precisely evaluate the network’s explanation ability.

Classification Metrics

For Camelyon16 patch-based dataset: We used Accuracy (Acc) and F1-score defined in Eqs. (4.5)-(4.6), as well as Area under the ROC curve (AUC) to evaluate the classification performance of our model and other methods.

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{4.5}$$

$$F1 - score = \frac{2TP}{2TP + FP + FN} \tag{4.6}$$

For BreakHis dataset: Two classification evaluation metrics were used on this dataset, i.e., patient recognition rate (PRR) and image recognition rate (IRR), detailed in Chapter 3.

Explainability

In general, there are mainly two folds of metrics used for evaluating networks’ performance on explanation. One is ground truth-independent metrics, such as Average drop%, % increase in confidence, and Win% in (Chattopadhay et al., 2018; desai and Ramaswamy, 2020), as well as Deletion and Insertion in (Petsiuk et al., 2021; Wang et al., 2020). Another is ground truth-based metrics, such as Dice index (Belharbi et al., 2021) and Precision (Huang and Chung, 2019; Li et al., 2021b).

However, All the above metrics have uncertainties in evaluating the ability of networks’ explanation in clinical settings. For instance, the ground truth-independent metrics calculate the effects of the explanation map on its network without considering the accuracy of target regions in the map that are important for human understanding in clinical diagnosis. Furthermore, Win% is only suitable for comparisons between two methods, and the calculation of Deletion and Insertion requires many forward passes of the network, which needs plenty of workloads. Dice index and Precision only consider the explanation that obeys human understanding while ignoring the evaluation of the network’s intrinsic explainability. Therefore, apart from the generally used Dice index in terms of both foreground ($Dice^{+}$) and background ($Dice^{-}$) regions, we also defined a new metric that accounts for both factors mentioned above to evaluate the explanation performance of the methods in this paper. Note that the formulation of $Dice$ is calculated as:

$$Dice = \frac{2\,|\mathrm{P} \cap \mathrm{G}|}{|\mathrm{P}| + |\mathrm{G}|} \tag{4.7}$$

where P and G are the binary target (foreground for $Dice^{+}$ and background for $Dice^{-}$) masks of the prediction and ground truth, respectively, and $|\cdot|$ the cardinality of a set.

Theoretically, A good enough category-corresponded explanation map should cover most of the target-relevant regions (that contribute to the category probability) in an image. Hence, when inputting the target region specified by a better explanation map instead of the whole image, the drop in the model's output confidence score is expected to be lower. That leads to a metric for the evaluation of explanation map:

$$AvgDrop\% = \frac{1}{N}\sum_{i=1}^{N}\frac{\max\left(0, O_i^c - X_i^c\right)}{O_i^c} \tag{4.8}$$

where N is the number of all the input images, O_i^c is the predicted confidence score for class c with original input image i, and X_i^c is the predicted confidence score for class c with only the explanation map region in image i as the input. The lower the $AvgDrop\%$ (i.e., Average drop%), the better the evaluation map. Yet, it has a fatal problem. If an explanation map covers almost the entire input image that includes the background, the $AvgDrop\%$ will approach zero, while such explanation map is meaningless.

On the other hand, Precision is an evaluation metric that calculates the rate of the correctly predicted region to the whole prediction, defined as:

$$Precision = \frac{|\mathrm{P} \cap \mathrm{G}|}{|\mathrm{P}|} \tag{4.9}$$

which decreases with the presence of more regions irrelevant to the target (i.e., more false positives or false negatives) in the explanation map. It also has a crucial problem. When the target region in an explanation map is inside the ground truth, no matter how tiny this target region is, Precision is equal to 1, while this explanation map is problematic.

In short, when the predicted target region is inside the ground truth and decreases in size (Precision is constant 1, i.e., the best), the $AvgDrop\%$ becomes worse (i.e., higher) due to the loss of relevant information that contributes to the decision. While the Precision deteriorates with the predicted target region tending to cover all the input images (including irrelevant pixels), the $AvgDrop\%$ approaches zero, i.e., the best. As we mentioned before, a clinically reliable and logical explanation map must satisfy the constraint that target regions indicated by a clinically human-understandable visual explanation map should highlight the lesion tissue as comprehensively as possible and have fewer false positives. Neither of these two metrics can define a explanation map satisfying that constraint. However, the predicted target regions in an explanation map meet the constraint when both the $AvgDrop\%$ and Precision have relatively optimal values. Therefore, we defined our new explanation metric Explainability as:

$$Explainability = (1 - AvgDrop\%) \times Precision \tag{4.10}$$

which can solve the issues of $AvgDrop\%$ and Precision since it increases when both metrics obtain better values (i.e., the explanation map is more clinically logical and understandable) and decreases otherwise. Thus, Explainability is more suitable for the evaluation of clinical explanation.

Furthermore, the confusion matrix over pixels, mean pixel accuracy (MPA), and mean pixel error (MPE, i.e., $1 - MPA$) were calculated to evaluate how well the model is in pixel-level classification.

$$MPA = \frac{1}{k} \sum_{i=0}^{k-1} \frac{p_{ii}}{\sum_{j=0}^{k-1} p_{ij}} \tag{4.11}$$

where $k = 2$ is the number of pixel categories, i.e., 0 and 1. p_{ii} is the number of correctly classified pixels with category i, and p_{ij} is the number of pixels classified as j while their true category is i.

4.3.2 Implementation Details

All the methods evaluated on the Camelyon16 patch-based dataset, including ours, were trained using the Adam optimizer with $\beta 1 = 0.9$, $\beta 2 = 0.999$, a weight decay of 5×10^{-4}, and a batch size of 8 for 512×512 patches except the ones based on MIL of which the batch size is 1. The learning rates for all the fine-tune layers and other layers without transferred parameters were initialized with 10^{-5} and 10^{-3}, respectively. All the methods were trained using image-level annotations and had no access to pixel-wise supervision. We used the binary cross-entropy loss to update the parameters in our model and followed the settings in the original works for other methods, including most of their hyper-parameters. As for the hyper-parameters that were not specified in the original works, we tuned them empirically through validation. For example, we set $kmax = 0.1$, $kmax = None$, class-related modalities to 4, and $\alpha = 0.7$ for WILDCAT. For CELNet-CELM, we set $\alpha = 0.8$ and $\beta = 0.2$. We trained all the methods for 50 epochs to find the best convergence. Training set was augmented by random horizontal and vertical flips with a probability of 0.5, random color jittering ($brightness = 0.5$, $contrast = 0.5$, $saturation = 0.5$ and $hue = 0.05$), and random rotation with an angle in $\{0°, 90°, 180°, 270°\}$ before being fed into all the networks. Particularly, for our ExplaCNet, we equally cropped each 512×512 input image into four patches with the size of 256×256 that compose a bag of instances before the augmentation.

For evaluation of the BreakHis dataset, we followed the same experimental protocol proposed in (Spanhol et al., 2016b), where the dataset was divided into a training set (70%) and a testing set (30%). The patients used to build the training set were not used for the testing set. We also trained all the networks for five trials and calculated the average value as the final result. Each input image with the size of 700×460 was overlap-cropped to six 256×256 patches before the augmentation for our model, and we resized the input to 456×456 pixels for the training of VGG16. The rest of the settings were the same as those adopted for Camelyon16 patch-based dataset.

4.4 Results and Discussion

4.4.1 Comparison in Performance of Explanation

We first compared the explanation performance of our ExplaCNet with many state-of-the-art explainable methods. They are Grad-CAM (Selvaraju et al., 2017a), Grad-CAM++ (Chattopadhay et al., 2018), CELNet-CELM (Huang and Chung, 2019), Ablation-CAM (desai and Ramaswamy, 2020), Score-CAM (Wang et al., 2020), WILD-CAT (Durand et al., 2017), ACoL (Zhang et al., 2018a), A-MIL (Ilse et al., 2018b; Patil et al., 2019), and Ciga and Martel (2021). Furthermore, we set two special explanation maps with all pixels equal to 0 (refereed as method All-zeros) and 1 (All-ones), respectively, to calculate the low baseline of explanation metrics, below which explainability will become unreliable. The All-zeros considers that all the input images are normal (or benign), while the All-ones identifies all the inputs as metastatic (or malignant). Note that we chose VGG16 as the backbone for all explainable methods except the backbone-fixed models because this makes more sense for the comparison with our ExplaCNet.

Table 4.1 provides the quantitative comparison results of the above methods on Camelyon16 patch-based dataset. It is seen that our ExplaCNet achieves the best performance in all explanation metrics compared to other explainable methods, with nearly 2.92% improvement in Dice^{+}, 2.38% improvement in Dice^{-}, and 5.55% improvement in Explainability against the second-best results, i.e., Dice^{+} of Score-CAM and Dice^{-} and Explainability of Grad-CAM. An important observation is that although most state-of-the-art methods can identify the tumorous regions with a Dice^{+} higher than the low baseline of 59.44% in the All-ones, they most perform poor in recognizing normal tissues with a Dice^{-} lower than the low baseline of 83.75% in the All-zeros. We can observe a similar result in Fig. 4.3, which gives the confusion matrix over all the pixels of the testing set and the MPA versus MPE. It shows that most state-of-the-art methods struggle to provide a good balance between true positives and true negatives. As can be seen in the visual comparison results in Fig. 4.4 and Fig. 4.5, these methods yield overdiagnosis (i.e., false positives). Thus, they have a weaker reliability in classification, although their Explainability is relatively higher than the low baseline of 54.89% provided by the All-zeros method. Our ExplaCNet, however, obtains a much better balance between true positives and true negatives and is able to identify both tumor and normal tissues with a higher Dice and MPA (a lower MPE), which leads to a higher Explainability, suggesting that our ExplaCNet produces more reliable classification results.

The visual comparison of all the methods depicted in Fig. 4.4 and Fig. 4.5 shows the explanation maps with metastatic regions marked in red, which provide evidence of the network decision. From the explanation maps of five metastatic examples (Fig. 4.4), we can see that metastatic regions indicated by our ExplaCNet are closer to the ground truth, while other methods tend to either yield overdiagnosis or produce more false negatives. In the visual explanation of five normal examples (Fig. 4.5), the ExplaCNet performs clearly better in identifying normal tissues and is more prone to correctly recognizing tumor-free input. Existing explainable methods, however, more or less produce false positives, i.e., identifying many normal tissues as metastatic,

Table 4.1: Comparison results (in %) of explanation and classification performance on Camelyon16 patch-based test set. The best performance is indicated in bold.

Methods	Explanation metrics			Classification metrics		
	$Dice^{+}$	$Dice^{-}$	Explainability	Acc	AUC	F1-score
All-ones (Lower-bound)	59.44 ± 0.00	0.00 ± 0.00	45.11 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	66.67 ± 0.00
All-zeros (Lower-bound)	0.00 ± 0.00	83.75 ± 0.00	54.89 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	0.00 ± 0.00
WILDCAT (Durand et al., 2017)	66.69 ± 1.87	78.78 ± 5.61	67.18 ± 3.50	**98.83 ± 0.33**	**99.88 ± 0.01**	**98.83 ± 0.33**
Grad-CAM (Selvaraju et al., 2017a)	64.31 ± 1.81	84.52 ± 1.38	70.42 ± 3.26	98.79 ± 0.18	99.72 ± 0.08	98.77 ± 0.19
ACoL (Zhang et al., 2018a)	67.43 ± 1.78	35.99 ± 6.18	64.20 ± 3.66	98.24 ± 0.43	99.61 ± 0.09	98.24 ± 0.42
Grad-CAM++ (Chattopadhay et al., 2018)	65.09 ± 1.85	80.41 ± 1.24	68.66 ± 3.38	98.79 ± 0.18	99.72 ± 0.08	98.77 ± 0.19
A-MIL (Ilse et al., 2018b; Patil et al., 2019)	50.68 ± 7.49	67.48 ± 11.66	59.18 ± 2.85	97.25 ± 0.33	98.88 ± 0.10	97.18 ± 0.35
CELNet-CELM (Huang and Chung, 2019)	58.30 ± 4.38	66.07 ± 4.60	53.49 ± 2.18	98.48 ± 0.13	99.13 ± 0.19	98.45 ± 0.13
Ablation-CAM (desai and Ramaswamy, 2020)	64.96 ± 0.73	70.19 ± 7.22	69.49 ± 2.49	98.79 ± 0.18	99.72 ± 0.08	98.77 ± 0.19
Score-CAM (Wang et al., 2020)	68.01 ± 0.72	77.71 ± 2.90	62.87 ± 3.49	98.79 ± 0.18	99.72 ± 0.08	98.77 ± 0.19
Ciga and Martel (Ciga and Martel, 2021)	58.66 ± 0.82	24.72 ± 5.60	48.26 ± 5.59	97.54 ± 0.80	99.52 ± 0.18	97.51 ± 0.81
ExplaCNet	**70.93 ± 1.16**	**86.90 ± 0.96**	**75.97 ± 2.32**	97.44 ± 0.28	99.61 ± 0.08	97.43 ± 0.27

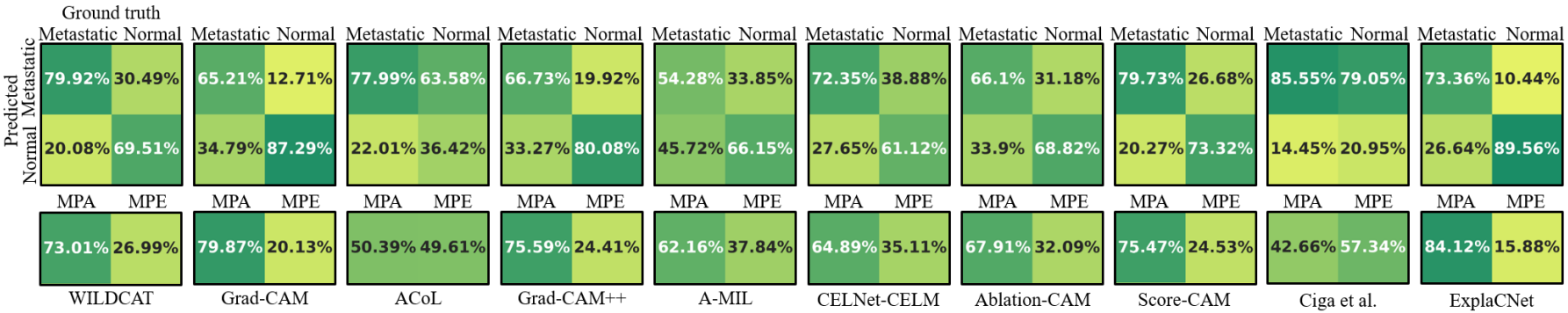

Figure 4.3: Top row: Confusion matrix over all pixels of Camelyon16 patch-based test set. Bottom row: MPA versus MPE.

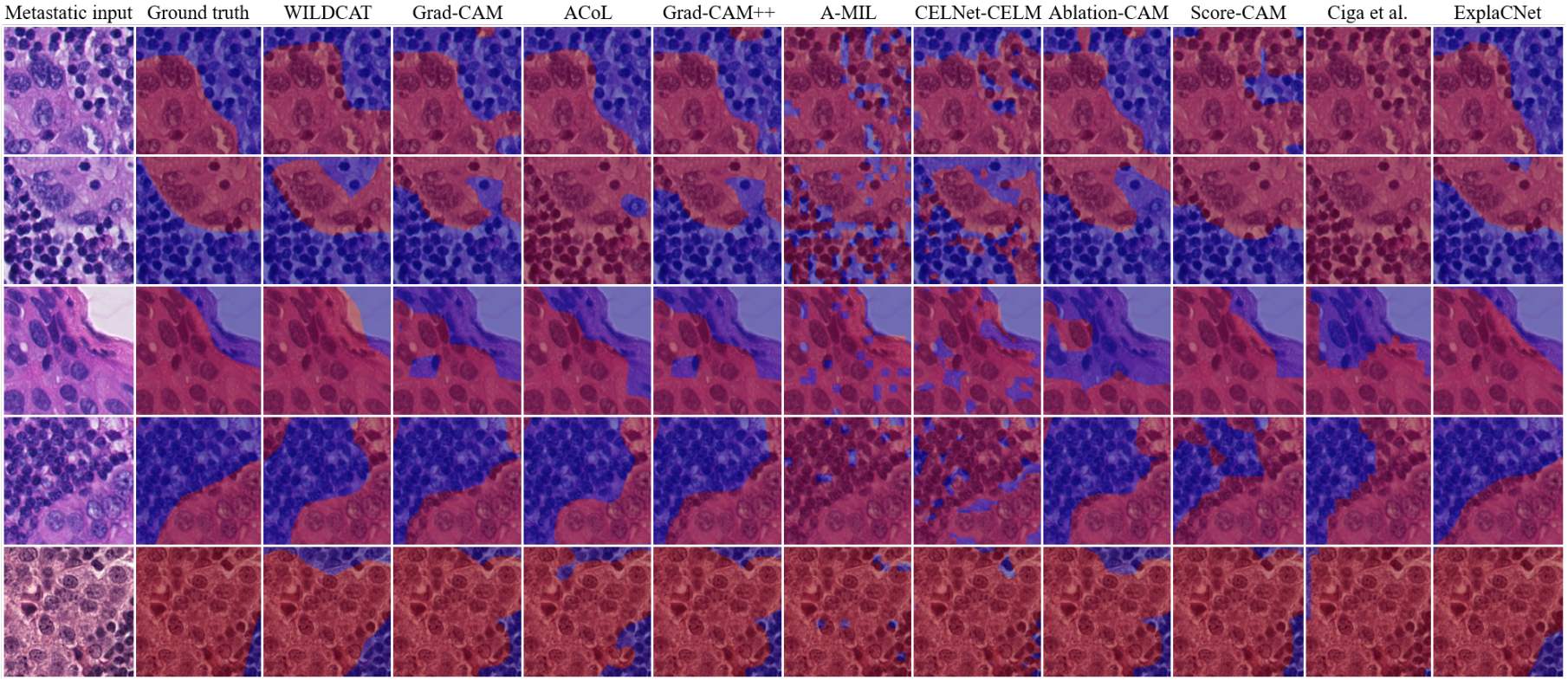

Figure 4.4: Examples of explanation maps for tumor input. Red and blue regions represent the presence and absence of metastatic tissues, respectively, which provide evidence for the classification decision.

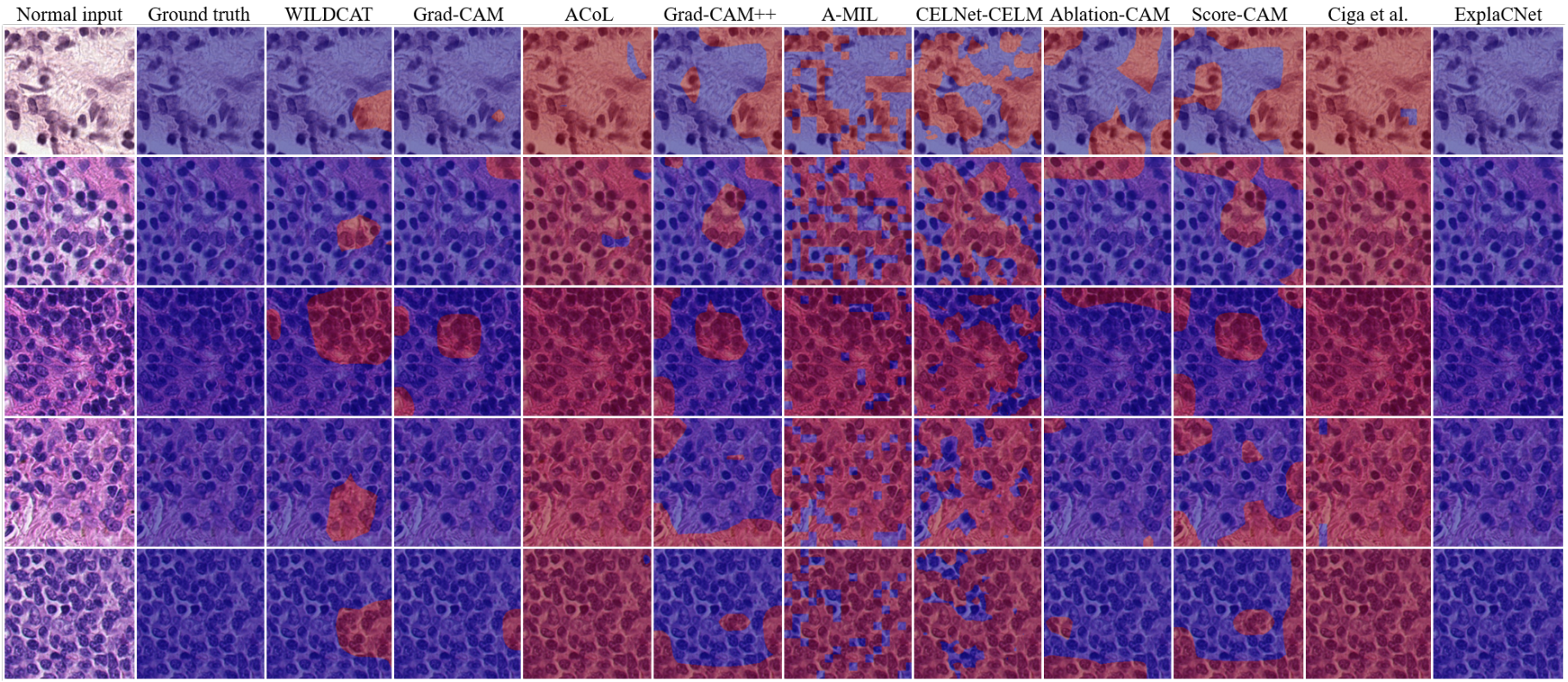

Figure 4.5: Examples of explanation maps for normal input. Red and blue regions represent the presence and absence of metastatic tissues, respectively, which provide evidence for the classification decision.

Table 4.2: Classification results (in %) on BreakHis test set. The best performance is indicated in bold.

Methods	40×		100×		200×		400×	
	PRR	IRR	PRR	IRR	PRR	IRR	PRR	IRR
Spanhol et al. (Spanhol et al., 2016a)	90.00 ± 6.70	85.60 ± 4.80	88.40 ± 4.80	83.50 ± 3.90	84.60 ± 4.20	82.70 ± 1.70	86.10 ± 6.20	80.70 ± 2.90
Bayramoglu et al. (Bayramoglu et al., 2016b)	83.08 ± 2.08	N/A	83.17 ± 3.51	N/A	84.63 ± 2.72	N/A	82.10 ± 4.42	N/A
Song et al. (Song et al., 2017)	90.20 ± 3.20	87.70 ± 2.40	91.20 ± 4.40	87.60 ± 3.90	87.80 ± 5.30	86.50 ± 2.40	87.40 ± 7.20	83.90 ± 3.60
Sudharshan et al. (Sudharshan et al., 2019)	92.10 ± 5.90	87.80 ± 5.60	89.10 ± 5.20	85.60 ± 4.30	87.20 ± 4.30	80.80 ± 2.80	82.70 ± 3.00	82.90 ± 4.10
Zhu et al. (Zhu et al., 2019)	85.20 ± 2.60	85.70 ± 1.90	83.50 ± 3.80	84.20 ± 3.20	84.10 ± 1.40	84.90 ± 2.20	79.30 ± 2.70	80.10 ± 4.40
Qi et al. (Qi et al., 2019)	91.26 ± 0.39	89.29 ± 0.50	**93.10 ± 0.32**	**90.95 ± 0.72**	92.84 ± 0.69	**91.61 ± 0.26**	**92.30 ± 0.37**	90.36 ± 0.83
Gour et al. (Gour et al., 2020)	87.47 ± 3.22	87.40 ± 3.00	88.15 ± 2.97	87.26 ± 3.54	92.52 ± 2.84	91.15 ± 2.30	87.78 ± 2.46	86.27 ± 2.18
SMSE (Sun et al., 2021)	87.51 ± 4.07	N/A	89.12 ± 2.86	N/A	90.83 ± 3.31	N/A	87.10 ± 3.80	N/A
VGG16 (Simonyan and Zisserman, 2014b)	**94.80 ± 1.93**	**93.85 ± 2.16**	91.83 ± 1.38	90.51 ± 1.57	92.23 ± 0.30	90.61 ± 0.42	89.11 ± 3.23	90.66 ± 2.75
ExplaCNet	93.69 ± 1.70	92.58 ± 2.25	90.45 ± 1.81	89.32 ± 2.74	**93.0 ± 0.35**	91.56 ± 0.97	89.34 ± 1.39	**91.07 ± 1.56**

especially for ACoL, A-MIL, CELNet-CELM, and Ciga et al. In short, our ExplaCNet can determine the normal or tumor of an image based on the absence or presence of tumor tissues like pathologists, which makes the classification more reliable and trustworthy.

4.4.2 Comparison in Performance of Classification

The comparison results of our ExplaCNet with other methods in classification performance are provided in Table Table 4.1 (classification metrics column) and Table 4.2. Table 4.1 shows that WILDCAT achieves the best performance in all three classification evaluation metrics on Camelyon16 patch-based dataset. Most state-of-the-art methods have an accuracy higher than 98%. Although our ExplaCNet obtains a slightly lower accuracy of nearly 97.44%, it is still competitive against some methods such as A-MIL. For the performance on the BreakHis dataset, we can see from Table 4.2 that our ExplaCNet outperforms most of the other methods and yields a slight boost on magnification factors of 200× and 400× compared with VGG16, suggesting that our explainability-favored reinforcement keeps the good classification performance of the network. It is worth mentioning that the classification task in our ExplaCNet satisfied the crucial constraint that human-level explainable or reliable classification requires the tumor regions specified by an explanation map to be in accord with the ones defined by pathologists as closely as possible. That could drive network optimization to favor a compromise solution to enhance explainability instead of an optimal classification. In contrast, existing state-of-the-art methods result in better classification results on Camelyon16 patch-based dataset, but they fail to provide a reliable explanation to support their decision compared with our ExplaCNet.

4.4.3 Ablation Study

We conducted a series of comparison experiments over the Camelyon16 patch-based dataset to evaluate the effectiveness of the key components adopted in our ExplaCNet, i.e., MIL, WAP, and DMG. Specifically, A base network (BN) with the first five convolutional layers of VGG16 as the FEM, a 1×1 convolution layer as the decision map generator and a general global average pooling layer as the classifier was constructed as the baseline. BN was trained with the settings identical to those of ExplaCNet except the input that was identical to that of methods without MIL. We investigated the effects of the above components on our ExplaCNet by gradually adding MIL, WAP, and DMG to BN, which resulted in three structures, i.e., BN+MIL (BM), BM+WAP (BW), and BW+DMG (ExplaCNet).

Table 4.3 provides the quantitative results of the ablation study. It shows that the MIL strategy favorably boosts the performance of BM in terms of all classification metrics as well as Dice^- and Explainability while producing a drop in Dice^+ compared with BN. Similarly, compared to BM, the WAP can help to boost BMW's performance in terms of Dice^+ and Explainability with a slight drop in classification accuracy and a negligible decrease in Dice^-. Furthermore, with the addition of the DMG module, our ExplaCNet's performance is improved in terms of all the explanation metrics by a non-negligible degree compared to BMW, i.e., nearly 2.21% improvement in Dice^+,

Table 4.3: Ablation study (in %) on Camelyon16 patch-based test set. The best performance is indicated in bold.

Methods	Explanation metrics			Classification metrics		
	$Dice^{+}$	$Dice^{-}$	Explainability	Acc	AUC	F1-score
BN	64.88 ± 1.00	81.80 ± 0.51	64.83 ± 0.74	93.71 ± 0.54	97.86 ± 0.67	93.55 ± 0.50
BN+MIL (BM)	62.68 ± 1.07	84.69 ± 0.63	70.98 ± 1.56	**98.18 ± 0.46**	**99.66 ± 0.05**	**98.18 ± 0.45**
BM+WAP (BMW)	68.72 ± 1.59	84.07 ± 1.71	72.38 ± 3.64	97.38 ± 0.35	99.56 ± 0.09	97.36 ± 0.36
BMW+DMG (ExplaCNet)	**70.93 ± 1.16**	**86.90 ± 0.96**	**75.97 ± 2.32**	97.44 ± 0.28	99.61 ± 0.08	97.43 ± 0.27

nearly 2.83% improvement in Dice^{-}, and nearly 3.59% improvement in Explainability. Interestingly, the classification accuracy is also enhanced by a slight degree.

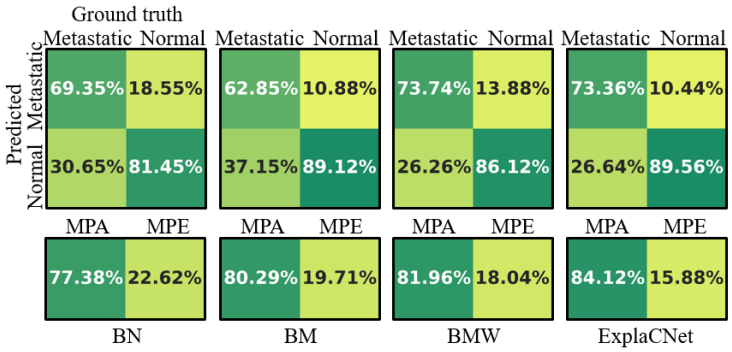

Figure 4.6: Top row: Confusion matrix over all pixels of Camelyon16 patch-based test set. Bottom row: MPA versus MPE.

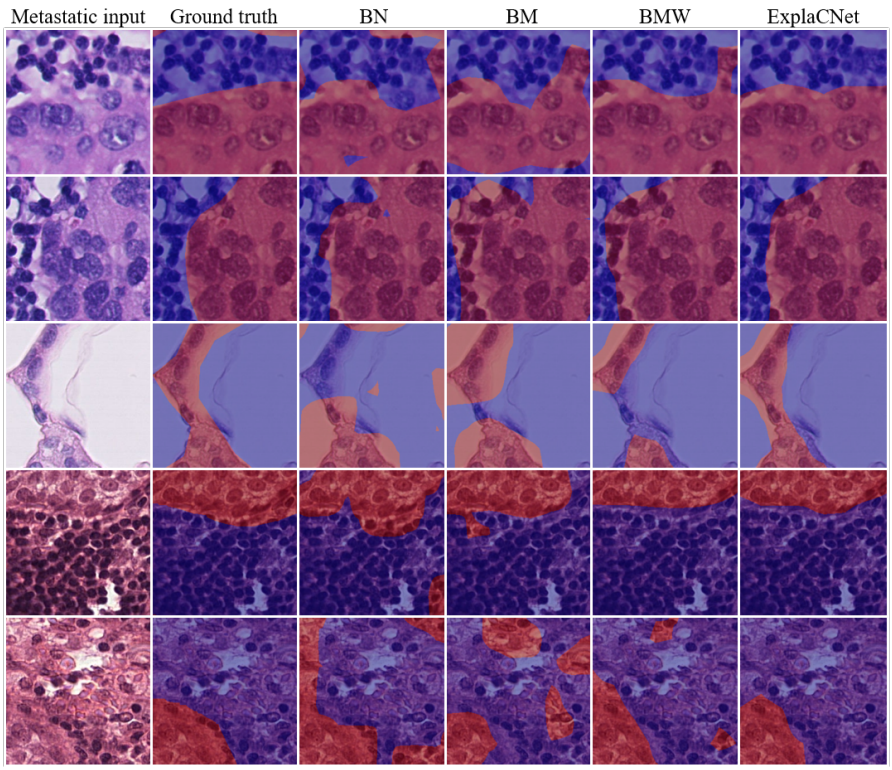

Figure 4.7: Examples of explanation maps for tumor input in ablation study. Red and blue regions represent the presence and absence of metastatic tissues, respectively, which provide evidence for the classification decision.

We can conclude from the above comparison results that the MIL strategy mainly helps to improve our ExplaCNet's ability to identify more normal tissues, thus increasing its classification performance. On the other hand, the WAP classifier contributes to encouraging our ExplaCNet to recognize more tumor regions. In particular, the DMG module allows our ExplaCNet to obtain a good compromise of the responsiveness to

tumor and normal tissues, thus optimizing the network's performance of identifying each of them. These inferences can be confirmed by Fig. 4.6. We can see that the BM has an improvement of 7.69% in true negative rate (TNR) and a deterioration of 6.5% in true positive rate (TPR) compared with BN. The BMW has an improvement of 10.89% in TPR and a deterioration of 3% in TNR compared with BM. Our ExplaCNet keeps the high TPR of BMW and the high TNR of BM. It is worth mentioning that every component increases our ExplaCNet's performance in terms of MPA, thus improving the Explainability.

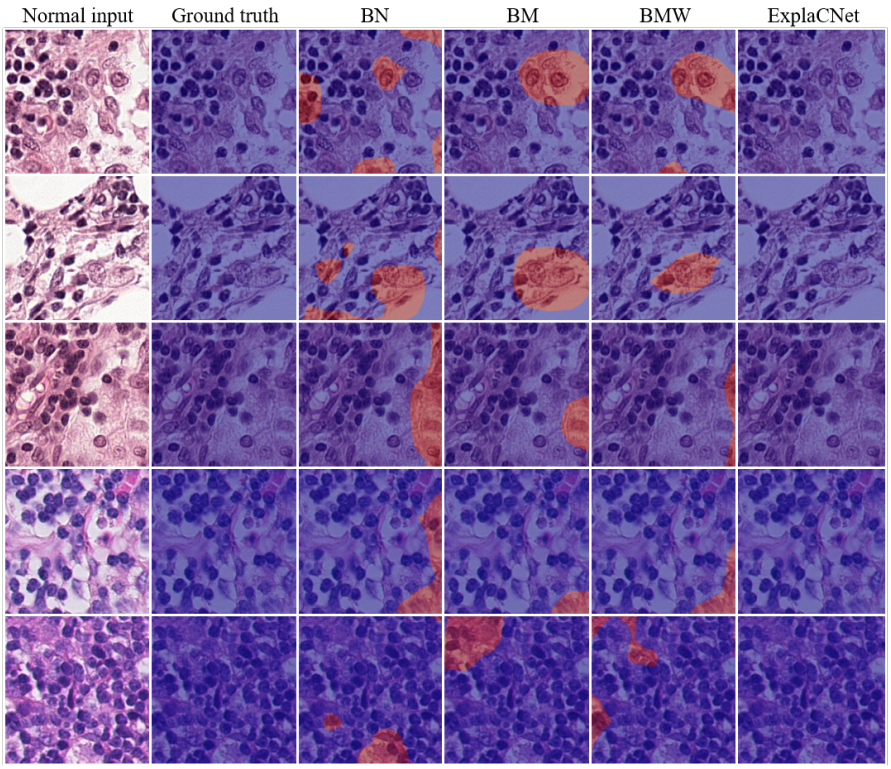

Figure 4.8: Examples of explanation maps for normal input in ablation study. Red and blue regions represent the presence and absence of metastatic tissues, respectively, which provide evidence for the classification decision.

The visual comparisons of the ablation study with tumor and normal inputs are shown in Fig. 4.7 and Fig. 4.8, respectively. It is observed in Fig. 4.7 that the tumor region indicated by the explanation map of the network gradually approaches the ground truth with the addition of our three key components. In Fig. 4.8, when all the components are embedded, our ExplaCNet becomes more reliable in mimicking human experts' diagnosis criterion (i.e., classifying histopathological images as normal or tumor based on the absence or presence of tumor tissues) to discriminate healthy images.

4.5 Conclusions

We proposed a novel WSL framework named ExplaCNet to achieve the explainable classification of histopathological breast cancer images in clinical diagnosis. The MIL strategy was used to crop large-size input into multiple instances to encourage ExplaCNet to identify more healthy tissues. A WAP classifier was developed to force ExplaCNet to recognize more tumor regions. Furthermore, we designed a DMG module on top of a deep feature extractor to optimize the network's attention on both tumor and normal pixels and generate reliable human-understandable explanation maps to support the classification results.

Extensive experimental results on the Camelyon16 patch-based dataset and BreakHis dataset demonstrated that our ExplaCNet outperformed state-of-the-art explainable models in visual explanation (i.e., classification reliability, which is of paramount importance for deep-learning methods to be deployed in real-world clinical settings) while keeping a competitive performance in classification. That superior performance makes our ExplaCNet more suitable for clinical diagnosis. Additionally, the ablation study verified the effectiveness of each key component of our network.

It is worth noting that our WSL framework can be readily ported to existing classification architectures to provide explainability and has great potential to achieve explainable whole-slide diagnosis. In the future, it would be interesting to put our method into practice and investigate the optimal trade-off between generating better human-understandable explanation maps and obtaining more accurate classification results in the clinical setting.

Chapter 5

EFFNet: Element-Wise Feature Fusion Network for Defect Detection of Display Panels

5.1 Introduction 126
5.2 Proposed Method 128
5.2.1 Defect Extraction Module 130
5.2.2 Feature Decoder 131
5.2.3 Element-Wise Feature Fusion Module 132
5.2.4 Loss Function 133
5.3 Experimental Settings 133
5.3.1 Implementation Details 133
5.3.2 Evaluation Metrics 134
5.4 Results and Discussion 134
5.4.1 Ablation Study 134
5.4.2 Comparison with Non-Deep-Learning Methods 135
5.4.3 Comparison with Deep-Learning Methods 138
5.4.4 Analysis of Failure Cases 149
5.5 Conclusion 150

5.1 Introduction

Display panels such as the thin-film-transistor liquid-crystal display (TFT-LCD) and organic light-emitting diode (OLED) are the main components of display products. However, the manufacturing process of these components is complicated and prone to suffer from different kinds of defects. The Array process is the first stage of the entire manufacture. Its commonly occurring defects have received extensive attention due to their adverse impact on the yield, life span, and function of display panels. Therefore, the detection of defects that arose from the Array process plays a vital role in the quality control and yield rate improvement of display panels in the actual production process. Generally, the defect inspection approach built on human vision is highly subjective, time-consuming, and labor-intensive. It hardly satisfies the increasing demand for real-time defect detection, i.e., high accuracy and speed (efficiency). The inception of automatic computer-vision-based defect detection techniques brings a solution to the above issue.

In the last few decades, a vast number of image processing methods have been applied to defect detection, such as Otsu (He and Sun, 2015; Ng, 2006; Yuan et al., 2015), Canny edge detector (Vasilic and Hocenski, 2006; Wang et al., 2018), Fourier transform (Tsai and Hung, 2005; Tsai et al., 2007; Barnes et al., 2013), Gaussian and Gabor filters (Mukherjee et al., 2006; Choi et al., 2014; Tong et al., 2016), optical flow (Tsai et al., 2011), and support vector machine (SVM) (Chu et al., 2017). For instance, defects in a glass substrate were detected by first establishing a straight-line interception histogram from the two-dimensional information of an image and then using Otsu criteria to find the best interception threshold (He and Sun, 2015). Fourier transform was used to remove the complex background and preserve local defects in TFT-LCD (Tsai et al., 2007). The recognition of five types of steel surface defects was achieved by four types of statistical features and enhanced twin SVM (Chu et al., 2017). A mura defect detection method (Jin et al., 2018) used the discrete cosine transform, the dual-γ piece-wise exponential transform, and Otsu was proposed for thin-film transistor liquid crystal display (TFT-LCD) panels. Detection of mura defects on liquid crystal display (LCD) under uneven brightness (Ma and Gong, 2019) was later studied through the Gabor filter, the background reconstruction algorithm based on the mura uniform light principle, and the gamma correction. However, the performances of these methods in real-time detection of defects with intricate textures still leave much to be desired. Intrinsically, the above methods mainly rely on handcrafted features and rules that are shallow in feature representation, such as gradient amplitude or local feature similarity. These properties make it difficult for them to effectively and integrally characterize target images, and they usually only perform well in combination with other techniques. Extrinsically, the majority of the features and the surrounding backgrounds of defects, such as the defects in display panels after the Array process, are rather complicated.

Recently, deep-learning-based methods have made a huge impact on the field of computer vision (Ouyang et al., 2016; Selvaraju et al., 2017b; Carbonneau et al., 2020; Isensee et al., 2021) due to their automatic feature learning and superior feature representation abilities. Convolutional neural networks (CNNs), such as VGG (Simonyan and Zisserman, 2014a) and ResNet (He et al., 2016a), have been the most groundbreaking addition. For image processing (Lei et al., 2018b; Li et al., 2019a; Le et al.,

2020; Ruan et al., 2020), CNNs have overwhelming performance in object recognition compared to traditional non-deep-learning methods. Therefore, CNNs have been consecutively applied to defect detection in industrial production. For example, a U-shaped network (Liu et al., 2020a) based on ResNet (U-ResNet) was proposed to accurately detect conductive particles after anisotropic conductive film (ACF) bonding in the TFT-LCD manufacturing process. Faster-RCNN (Hu and Wang, 2020) was improved by introducing ResNet50 with feature pyramid networks as the backbone and was used for the detection of printed circuit boards. An end-to-end trainable deep convolutional neural network, DeepCrack (Zou et al., 2018) built on SegNet (Badrinarayanan et al., 2017), was proposed for automatic crack detection. Encoder-decoder residual network (EDRNet) (Song et al., 2020) with the combination of deep supervision mechanism and fusion loss was developed to detect the surface defects of strip steel. Lee et al. (Lee et al., 2022) adopted VGG16 to study the classification of defects in TFT-LCD panels. They used several methods including integrated gradients, SmoothGrad, Decovent, guided backpropagation, and deep Talyor to achieve the post hoc analysis of classification. The same year, an end-to-end multi-task learning network architecture (Li and Li, 2021) that contains an encoder, a feature fusion module, a segmentation head, and a classification head was proposed for the defect detection of mobile phone light guide plate (LGP). Furthermore, a multi-category classification model (Chang et al., 2022) based on the convolutional neural network working with automatic optical inspection (AOI) was proposed for identifying defective pixels on the TFT-LCD panel. Yao et al. (Yao and Li, 2022) proposed an AYOLOv3-Tiny network in combination with an overlapping pooling spatial attention module (OSM) and a dilated convolution module (DCM) for the defect detection of LGPs. An improved RetinaNet that used ResNeXt50 as the backbone (Li and Wang, 2022) was proposed for LGP defects detection, which adopted a Ghost module to replace the 1×1 convolution in the lower half of the ResNeXt block to reduce the resource parameters and consumption.

Figure 5.1: Typical hard-to-recognize defects in display panels. (a) The weak defect in the light-colored background. (b) The strong defect in the dark-colored background. (c) The malformation defect with an ambiguous boundary. The red marks indicate the regions of defects.

Despite CNN models having made considerable progress in defect detection, the

problem of online real-time defect detection was still not comprehensively addressed, especially for the inspection of display panel production. First of all, deep learning is a data-driven technique and the performance of a CNN model is highly subjected to training data. Yet, it is both time-consuming and labor-intensive to obtain a training dataset of display panels with sufficient numbers and variety. On the other hand, the detection of defects in display panels faces many unsolved challenges: 1) low contrast between defect and surrounding background has two forms: weak defects in light-colored background (Fig. 5.1(a)) and strong defects in dark-colored background (Fig. 5.1(b)); 2) intricate and diversiform background noise in display panels; 3) malformation defects with ambiguous boundaries: these defects in display panels have an identical color with the surrounding defect-free area, which cannot be distinguished by intensity difference (Fig. 5.1(c)). These factors undoubtedly bring harsh requirements for the design of a fast and accurate CNN model.

Inspired by Attention U-Net (Schlemper et al., 2019), which shows an impressive performance in segmentation tasks even with a scarce amount of labeled training data, we developed an efficiency-favored encoder-decoder configuration of our ConfigNet to overcome the aforementioned problems and achieve real-world online defect detection of display panels with high accuracy. This configuration is named EFFNet, i.e., element-wise feature fusion network. Considering the difficulty of gathering a large enough dataset of display panel defects, a transfer learning strategy and data augmentation were used in our method.

The main contributions of this paper are listed as follows:

- A novel EFFNet model based on VGG16 and encoder-decoder architecture was proposed to address the problem of real-time defect detection of display panels with multi-class backgrounds.
- To the best of our knowledge, this is the first attempt to adopt a CNN model embedded with the additive attention mechanism to solve the problem of online defect detection of display panels after the Array process.
- We designed a transfer learning and fine-tune strategy for the feature extraction module. It is effective in speeding up network training and increasing detection performance.
- We developed a feature fusion module based on the element-wise addition of size-matched pyramid features. It highlights regions of interest (ROIs) on the same channel of feature maps between shallow and deep features, which exceedingly shortens detection time and improves the identification of defect pixels.

5.2 Proposed Method

The overall architecture of our EFFNet model, as depicted in Fig. 5.2, is composed of an encoder and a decoder. The encoder, i.e., defect extraction module, is to extract defect features from complex background layer by layer and form a spatial feature pyramid structure that contains the semantic information of defects at different levels. The decoder embedded with element-wise feature fusion modules, on the other hand,

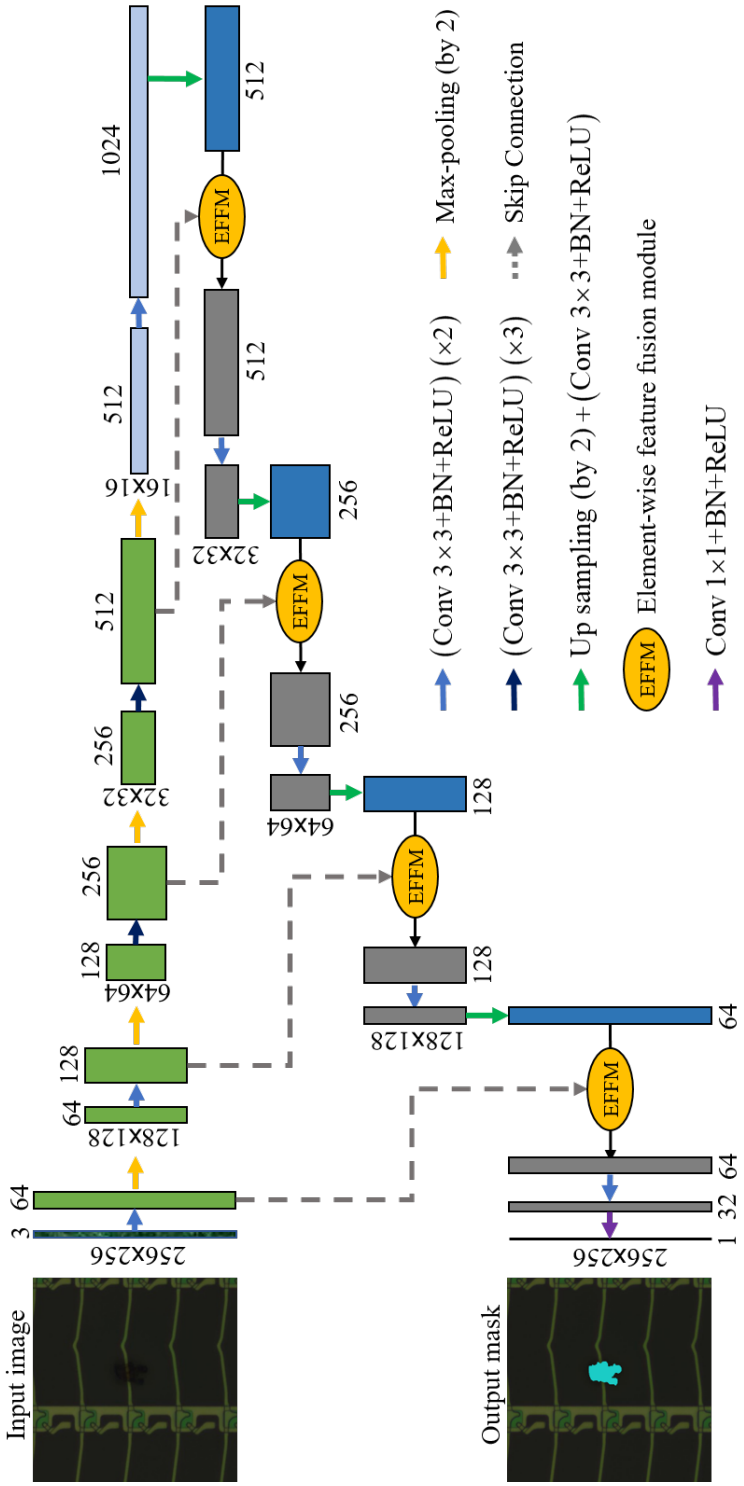

Figure 5.2: An illustration of the proposed EFFNet model.

gradually integrates the defect information from the encoded features. After the above processes, the segmentation results that indicate the localization of defects in display panels are obtained.

5.2.1 Defect Extraction Module

To mitigate small dataset overfitting and shorten training time, we adopted the convolutional blocks (ConvBlocks) of VGG-16(Simonyan and Zisserman, 2014a) model pre-trained by ImageNet dataset as our defect extraction module. The module extracts a set of hierarchical features with different scales that encode multilevel contextual information from the input image. The detailed structure of the module that contains five ConvBlocks is given in Table 5.1. Each block consists of several 3×3 convolutional layers and one 2×2 max-pooling layer with a stride of 2 at the end (except the last block). Considering that the wider the network is, the more interactive cross-channel information the network will have, we replaced the last ConvBlock of VGG-16 with two convolutional layers of 512 dimensions and 1024 dimensions to increase the network width and enhance the detection performance. Here, we use $\mathrm{E}_l(\cdot)$ to denotes the lth ConvBlock operation of the extraction module for simplicity, where $l \in \{1, 2, 3, 4, 5\}$. Thus, the extracted features f_l of the lth ConvBlock of the module are calculated as

$$f_l = \mathrm{E}_l(f_{l-1}) \tag{5.1}$$

where f_{l-1} is the output features of the $(l-1)$th ConvBlock and f_0 (i.e., $l = 1$) is the input image. $\mathrm{E}_l(\cdot)$ changes with l according to Table 5.1.

Table 5.1: Defect Extraction Module.

ConvBlock	Layer
1	$[\text{Conv } 3\times3 + \text{BN} + \text{ReLU}, \text{C} = 64] \times 2$ $\text{Maxpool } 2\times2$
2	$[\text{Conv } 3\times3 + \text{BN} + \text{ReLU}, \text{C} = 128] \times 2$ $\text{Maxpool } 2\times2$
3	$[\text{Conv } 3\times3 + \text{BN} + \text{ReLU}, \text{C} = 256] \times 3$ $\text{Maxpool } 2\times2$
4	$[\text{Conv } 3\times3 + \text{BN} + \text{ReLU}, \text{C} = 512] \times 3$ $\text{Maxpool } 2\times2$
5	$\text{Conv } 3\times3 + \text{BN} + \text{ReLU}, \text{C} = 512$ $\text{Conv } 3\times3 + \text{BN} + \text{ReLU}, \text{C} = 1024$

In particular, the max-pooling used in our network is to reduce the resolution of input images and expand the receptive field of the network while keeping translation invariance over small spatial shifts. Yet, continuous reduction of the resolution of feature maps would bring the loss of information related to boundary detail to some extent. Therefore, all the max-pooling indices were saved and used in the subsequent corresponding up-sampling layers to keep the integrality of defect representation. Furthermore, batch normalization was added after each convolutional layer to accelerate

training and prevent over-fitting, followed by a rectified linear unit (ReLU) activation function to increase the nonlinearity. The feature channels were doubled every time after one block operation while the feature sizes were halved.

5.2.2 Feature Decoder

As described in previous subsection, f_l ($l \in \{1, 2, 3, 4, 5\}$) is the output features of the lth ConvBlock of the encoder. In our decoder, the feature maps obtained from the fifth ConvBlock (i.e., $l = 5$) were first up-sampled to twice their original size by using the stored max-pooling indices from the forth layer of the encoder. Then, a 3×3 convolution operation was added to increase local contextual information and reduce the feature channels by half. The mathematical representation is formulated as

$$x_1 = \sigma_1 \left(W_{3\times3} * U \left(f_5 \right) + b \right) \tag{5.2}$$

where x_1 refers to the up-sampled high-level features of the first decoder layer and σ_1 is the ReLU activation function. $W_{3\times3}$ and b represent the 3×3 convolution kernel and bias, respectively. "$*$" denotes convolution operator and $U(\cdot)$ the up-sample operation with saved max-pooling indices and down-sample rate of the corresponding encoding layer, respectively.

Afterward, an element-wise feature fusion module (EFFM) was developed to guide the skip connection between shallow and deep features, followed by two 3×3 convolutional layers that halve the feature channels. The mathematical representation is given by

$$f'_1 = \Phi \left(\wp \left(f_4, x_1 \right) \right) \tag{5.3}$$

where f'_1 and $\Phi(\cdot)$ refer to the output of the first decoder layer and the convolution operation of two 3×3 convolutional layers, respectively. $\wp(\cdot)$ denotes the EFFM operator.

Similarly, the outputs of the other decoder layers are calculated as

$$f'_m = \Phi \left(\wp \left(\sigma_1 \left(W'_{3\times3} * U \left(f'_{m-1} \right) + b \right), f_{5-m} \right) \right) \tag{5.4}$$

where f'_m ($m \in \{2, 3, 4\}$) represents the output of the ith decoder layer. $W'_{3\times3}$ designates the 3×3 convolution kernel that remains feature channels.

Finally, the segmentation result Y is computed as

$$Y = \sigma_2 \left(W_{1\times1} * f'_4 + b \right) \tag{5.5}$$

where

$$\sigma_2 \left(z \right) = \frac{1}{1 + e^{-z}} \tag{5.6}$$

is the sigmoid activation function for normalizing the segmentation score. $W_{1\times1}$ denotes the 1×1 convolution kernel for decoupling the features and mapping them to lower-dimensional space.

5.2.3 Element-Wise Feature Fusion Module

The EFFM (i.e., element-wise feature fusion module) with an additive attention mechanism was developed with the inspiration of Attention U-Net (Schlemper et al., 2019) and is presented in Fig. 5.3. It used high-level features that have relatively more global semantic information as the gating signal to drive the network to focus more on the target pixels. Particularly, the up-sampled feature map $x_{l'}$ ($l' \in \{1, 2, 3, 4\}$) from the decoder layer l' and the feature map f_l ($l < 5$) from the encoder layer l were first halved in channels by $W_{1\times1}$ convolution operation for spatial information extraction. The halved deep-level and shallow-level features were then combined through element-wise addition, followed by the ReLU activation function. The resulting features were passed to a $W_{1\times1}$ convolution layer again to generate a single-channel image-grid attention map. Finally, the Sigmoid function σ_2 was adopted to obtain the attention coefficients ξ_l. Mathematically, the attention coefficients for filtering the output features from the encoder layer l are computed as

$$\xi_l = \sigma_2\left(W_{1\times1} * (\sigma_1\left(W_{1\times1} * f_l \oplus W_{1\times1} * x_{l'} + b_1\right)) + b_2\right) \tag{5.7}$$

where $\oplus$ denotes the element-wise addition operator. b_1 and b_2 refer to biases.

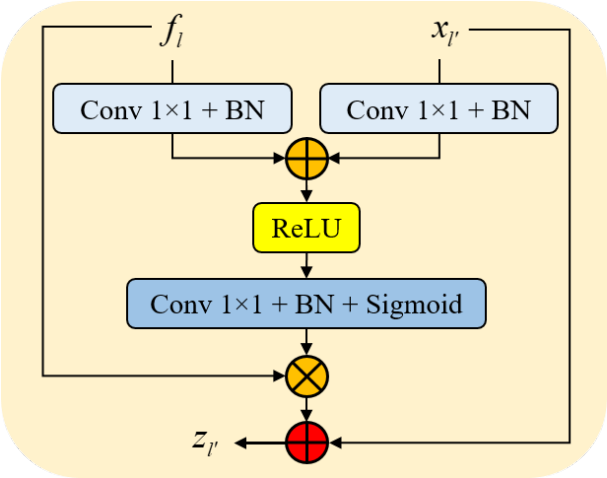

Figure 5.3: The architecture of element-wise feature fusion module.

The obtained attention coefficients were then used to filter the shallow features from the encoder by multiplication. Thereafter, the results were fused with the deep features that were expected to pass to the next layer of the decoder. Here, we used element-wise addition to achieve the fusion operation instead of the channel-wise concatenation that was adopted in Attention U-Net. The key reason is that the former forces the feature fusion module to learn to underline the target regions (i.e., ROIs) on the same channel of feature maps between shallow and deep features during the training, which avoids introducing extra parameters to the subsequent layers and thus significantly reducing computational complexity. Additionally, reusing coarse information through this improvement in skip connection can facilitate the network to learn more relevant information that improves detection performance. Mathematically,

the EFFM output $z_{l'}$ of the l'th decoder layer can be formulated as

$$\begin{aligned} z_{l'} &= \wp\left(f_l, x_{l'}\right) \\ &= \xi_l \otimes f_l \oplus x_{l'} \end{aligned} \tag{5.8}$$

where $\otimes$ refers to the element-wise product operator.

It is worth mentioning that the attention coefficient map can be learned according to different levels of defect information during the training since the attention works in different layers of the decoder. Furthermore, the attention coefficients can filter the activations during both backward and forward passes, as formulated in Eq. 5.9

$$\begin{aligned} \frac{\partial z_{l'}^i}{\partial W} &= \frac{\partial\left(\xi_l^i \cdot f_l^i + x_{l'}^i\right)}{\partial W} \\ &= \xi_l^i \frac{\partial f_l^i}{\partial W} + \frac{\partial \xi_l^i}{\partial W} f_l^i + \frac{\partial x_{l'}^i}{\partial W} \end{aligned} \tag{5.9}$$

where $\xi_l^i \frac{\partial f_l^i}{\partial W}$ is scaled with ξ_l^i, which suppresses gradients derived from regions irrelevant to the target while encouraging the network to gradually learn more relevant features with the increase of the decoder layer.

5.2.4 Loss Function

We updated the parameters of the proposed network by minimizing a pixel-wise loss function (binary cross-entropy)

$$L(g, p) = -\frac{1}{K} \sum_{i=1}^{K} \left[g_i \cdot \log\left(p_i\right) + \left(1 - g_i\right) \log\left(1 - p_i\right)\right] \tag{5.10}$$

where K is the batch size. g and $g_i \in \{0, 1\}$ are the ground truth mask and the ground truth of the ith image, respectively. p and $p_i \in [0, 1]$ are the network prediction and the predicted probability of the ith image, respectively.

5.3 Experimental Settings

5.3.1 Implementation Details

All the experiments were implemented on the Pytorch framework using a single NVIDIA GTX 2080 Ti GPU (with 11G memory) on windows 10. The initialization parameters of the first four ConvBlocks of our network were loaded from the pre-trained VGG-16 network, while the weights of other layers were initialized with the "Kaiming" initializer. For fine-tuning, the learning rates of the pre-trained layers were set to be 10^{-5}, and 10^{-4} was chosen for other layers. We used an Adam optimizer with a batch size of 12 images randomly cropped to 256×256 pixels from input images of 288×288 pixels to train our network.

5.3.2 Evaluation Metrics

To evaluate the performance of our EFFNet model, the mIoU (Long et al., 2015), MPA (mean pixel accuracy, detailed in Chapter 4), F1-Measure over pixels (a.k.a. Dice index, detailed in Chapter 4), and the precision-recall (P-R) curve were calculated between the ground truth and the predicted segmentation result.

We also defined threshold-based detection rate (TDR) expressed as

$$\mathrm{TDR}_{\eta} = \frac{N_{\mathrm{mIoU}\geq\eta}}{N_{\mathrm{img}}} \times 100\% \tag{5.11}$$

where η designates the threshold of the detection rate, $N_{\mathrm{mIoU}\geq\eta}$ is the number of images whose mIoU is over η, and N_{img} denotes the number of input images. Notably, the acceptable threshold of mIoU is 0.5 in our task. Therefore, we chose 0.5 as the minimum mIoU threshold and took 0.05 as the interval to compute the TDR-mIoU curve to further evaluate the methods.

5.4 Results and Discussion

5.4.1 Ablation Study

This study aimed to verify the effectiveness of EFFM component and transfer learning strategy used in our method.

Effects of EFFM

To evaluate the proposed EFFM used in our method, we implemented a series of comparative experiments: 1) simple encoder-decoder architecture without skip connection: we named it as BaseNet (BN) for simplicity; 2) the BaseNet with skip connection based on simple element-wise addition: named as BN+EA; 3) the BaseNet with feature fusion module based on channel-wise concatenation (brought from Attention U-Net): simplified as BN+CFFM; 4) the BaseNet with the proposed EFFM: BN+EFFM (ours).

Table 5.2: Evaluation results of ablation study in the effects of EFFM.

Method	mIoU	MPA	F1-Measure	Speed (fps/s)
BN	0.8108	0.8904	0.7311	**184**
BN+EA	0.8282	**0.9055**	0.7613	178
BN+CFFM	0.8364	0.9023	0.7775	124
BN+EFFM (ours)	**0.8377**	0.8941	**0.7867**	159

The evaluation results are provided in Table 5.2. Values in bold are the best results and this marking is applied to all the tables in this paper. It is shown that the EFFM favorably boosts the performance of our method in terms of mIoU and F1-Measure compared to simple encoder-decoder architecture BN and BN+EA. Moreover, compared with the channel-wise concatenation from Attention U-Net (i.e., BN+CFFM), the element-wise addition fusion strategy in EFFM not only significantly increases the

network's detection efficiency (by 35 fps/s) but also improves its performance in mIoU and F1-Measure, which denotes the effectiveness of our refinement. The small decrease in MPA, as we conjecture, is due to the fact that the EFFM prefers to improve the mIoU and F1-Measure in a way that activates more target regions rather than trying to discredit more irrelevant pixels, which to some extent leads to an increase in false positives that is tolerable and negligible in our task.

Transfer Learning Strategy Analysis

Since allowing different layers of the pre-trained feature extraction module to be fine-tuned may lead to the variable performance of our network in defect detection, we compared five potential layer-freezing schemes, i.e., freezing all the pre-trained layers (Encoder-F4), freezing the first three ConvBlocks (Encoder-F3), freezing the first and second ConvBlocks (Encoder-F2), freezing the first ConvBlock (Encoder-F1), and not freezing any layer (Encoder-F0, namely ours), to verify the effectiveness of the selected transfer learning strategy.

Table 5.3: Evaluation results of our encoder with five transfer learning strategies.

Method	mIoU	MPA	F1-Measure
Encoder-F4	0.8347	0.8884	0.7829
Encoder-F3	0.8299	0.8901	0.7684
Encoder-F2	0.836	0.8867	0.781
Encoder-F1	0.8297	0.8852	0.7718
Encoder-F0 (ours)	**0.8377**	**0.8941**	**0.7867**

It can be seen from the comparison in Table 5.3 that, among all the five transfer learning strategies, the performance of Encoder-F0 is the best. It indicates that, in our task, freezing layers of the pre-trained extraction module will prevent the network from learning more information relevant to the increasing of segmentation performance. Hence, Encoder-F0 was selected for our method.

5.4.2 Comparison with Non-Deep-Learning Methods

To better demonstrate the effectiveness and extraordinary performance of our EFFNet model in defect detection of display panels, we compared it with four typical non-deep-learning segmentation techniques. They are Canny edge detector (Canny) in (Vasilic and Hocenski, 2006), adaptive binarization method (Otsu) in (Ng, 2006), seeded region growing (SRG) in (Pohle and Toennies, 2001), and morphological geodesic active contour (M-GAC) algorithm in (Marquez-Neila et al., 2013), respectively. Theoretically, the techniques mentioned above require a threshold or initialization to complete the segmentation of objects. To fully explore the upper limit of these methods (except Canny) in defect detection in our task, we added another comparative experiment group that uses the positions and pixel values of the ground truth as the initialization and threshold.

The quantitative evaluation results of the comparison are given in Table 5.4. It can be observed that all four traditional image processing methods perform poorly, with all

the mIoU values lower than 0.5. We surmise the main cause is that these methods fail to find suitable initialization or thresholds in the complex defect background. When we fix the initial location and threshold for the corresponding methods (marked by "*") according to the ground truth, the evaluation metrics are increased significantly. They are 0.2209 (SRG) to 0.3351 (Otsu) higher in mIoU, 0.2705 (Otsu) to 0.3505 (SRG) higher in MPA, and 0.3038 (Otsu) to 0.4793 (M-GAC) higher in F1-Measure. Although the highest mIoU, MPA, and F1-Measure among these methods are raised to 0.7018 (M-GAC), 0.8930 (SRG), and 0.5496 (M-GAC), respectively, they are still lower than that of our EFFNet model. In our case, the values are 0.8377 in mIoU, 0.8941 in MPA, and 0.7861 in F1-Measure, respectively. The histogram of the comparison between our method and the other three traditional techniques based on the ground truth is shown in Fig. 5.4. It is observed that the EFFNet exceedingly overshadows the non-deep-learning segmentation techniques (even with the ground truth information).

Table 5.4: Evaluation results of our method and other non-deep-learning methods.

Methods	mIoU	MPA	F1-Measure
Canny	0.4761	0.5085	0.0234
Otsu	0.2916	0.4746	0.0125
M-GAC	0.4555	0.547	0.0703
SRG	0.4226	0.5425	0.0302
Otsu*	0.6267	0.7451	0.3163
M-GAC*	0.7018	0.846	0.5496
SRG*	0.6435	0.893	0.417
Ours	**0.8377**	**0.8941**	**0.7867**

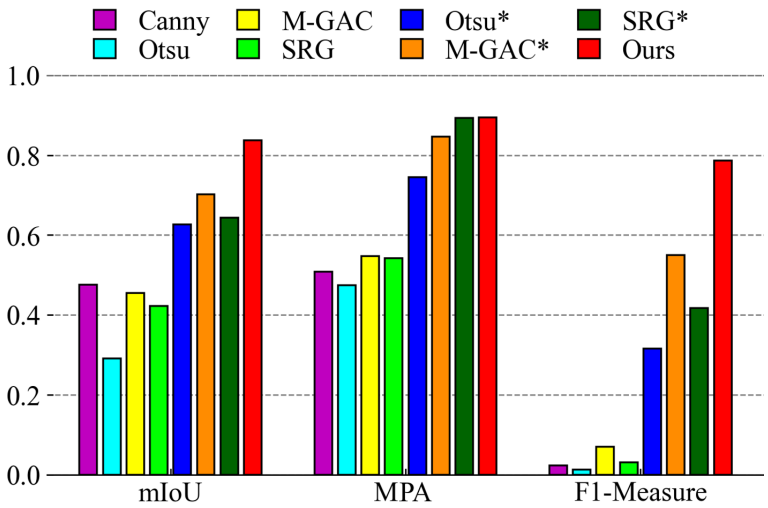

Figure 5.4: Histogram of the evaluation results of different methods.

A visual comparison of the proposed network with the three non-deep-learning methods based on the ground truth is depicted in Fig. 5.5. It shows that the traditional segmentation techniques, in many cases, either fail to detect the defects or have a high false-positive rate, even with the initialization information from the ground truth. For example, the results of Otsu* in class 1, class 4, and class 5, the results of M-GAC* and

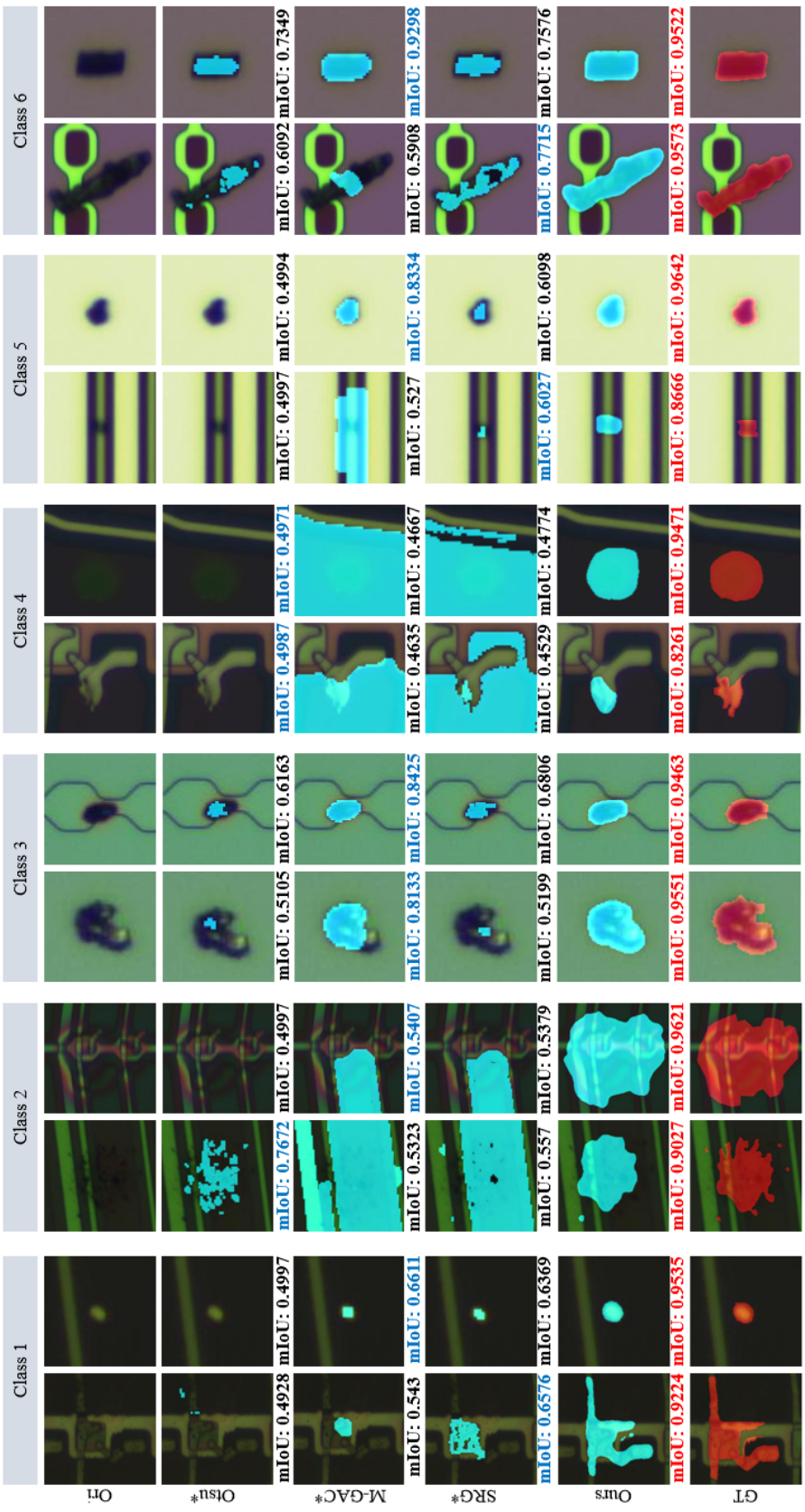

Figure 5.5: Comparison of the proposed method with traditional image processing methods. Ori denotes the original image, and GT means the ground truth. The numbers at the bottom of the images represent the corresponding mIoU values of the detection, where the values in red are the best, and the ones in blue are the second-best.

SRG* in class 2 and class 4. On the contrary, our EFFNet model is capable of finding defects in challenging cases with high accuracies, such as the low contrast cases in the first column of class 2 and the second column of class 4, the complex texture case in the second column of class 2, and the malformation cases in the first columns of class 1 and class 4, where the predictions are closer to the ground truth. These results are consistent with the previous quantitative analysis.

5.4.3 Comparison with Deep-Learning Methods

Meanwhile, we also compared our EFFNet with five state-of-the-art defect detection and object segmentation networks to further expound the superiority and better applicability of the proposed architecture in real-time defect detection of display panels, i.e., Attention U-Net (Schlemper et al., 2019), MultiResUNet (Ibtehaz and Rahman, 2020), U-ResNet (Liu et al., 2020a), DeepCrack (Zou et al., 2018), and EDRNet (Song et al., 2020).

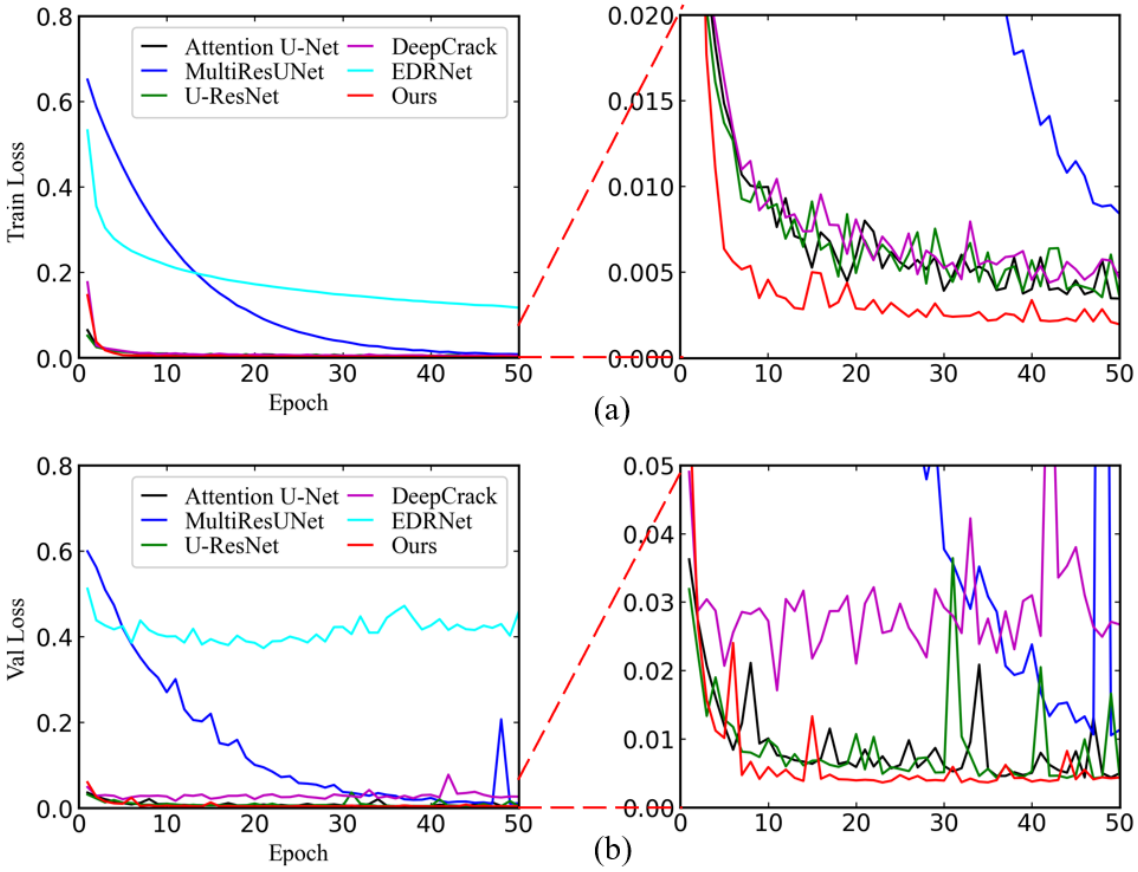

Figure 5.6: Learning curves of CNN-based methods. (a) Training loss. (b) Validation loss. The curves on the right are the enlargements of the corresponding areas of the left learning curves.

Fig. 5.6 gives the learning curves of our EFFNet model and the other five networks. It is worth noting that, unlike the other methods, the loss function of EDRNet is the addition of BCE loss, IOU loss, and SSIM loss Song et al. (2020). Thus, only the

convergence of the EDRNet is compared in this experiment. It is observed that the proposed model converges faster than the other methods due to the pre-trained layers in its encoder, and it has the lowest overall loss value and smoother loss curve when the loss is getting stable.

Table 5.5: Quantitative comparison of deep-learning methods.

Methods	mIoU	MPA	F1-Measure	Speed (fps/s)
Attention U-Net	0.807	0.8646	0.7239	111
MultiResUNet	0.7907	0.8565	0.6729	138
U-ResNet	0.8121	0.869	0.7292	83
DeepCrack	0.809	0.8689	0.7363	50
EDRNet	0.822	0.8697	0.7564	60
Ours	**0.8377**	**0.8941**	**0.7867**	**159**

The quantitative evaluation results of the deep-learning methods are provided in Table 5.5. Clearly, our EFFNet model outperforms other networks in terms of all three evaluation metrics. The EDRNet is the second-best model, compared to which our method is 1.57% higher in mIoU, 2.44% higher in MPA, and 3.03% higher in F1-Measure. In particular, we also investigate the processing time of the CNN-based techniques in defect detection with an input size of 256×256, and the results are presented in the last column of Table 5.5. As observed, our EFFNet model achieves the fastest speed, i.e., 159 fps/s. It is 2.5 times faster than that of EDRNet and 21 fps/s faster than that of the second-fastest model, i.e., MultiResUNet, while the latter has the worst performance in the three evaluation metrics. The TDR-mIoU and P-R curves in Fig. 5.7(a) and Fig. 5.7(b) also show the superior performance of our method, from which we can see that our EFFNet model has the biggest TDR over the whole mIoU threshold and the highest precision over a large range of thresholds. The histogram of the comparison between these methods is illustrated in Fig. 5.7(c).

Fig. 5.8 shows the visual comparison of our method with other state-of-the-art CNNs for the segmentation of twelve typical defective images. We can observe that, similar to the traditional image processing techniques, the other five state-of-the-art networks fail to detect defects in many challenging cases. For instance, MultiResUNet and U-ResNet fail to detect the tiny adhesive defect in the second column of class 1 and the malformation defect in the first column of class 4. Attention U-Net and DeepCrack fail to detect the low contrast part of the defect in the first column of class 5, which also happens to U-ResNet and EDRNet. For the image in the first column of class 3, all the other four networks (except EDRNet) mistakenly connect two tiny foreign objects as one large defect. Furthermore, all the other five networks fail to locate the malformation defect in the first column of class 6. Our EFFNet model, on the contrary, can detect the above defects with the segmentation results much closer to the corresponding ground truths. Notably, even though other methods can detect defects with high accuracy in many cases, their performance is still lower than that of the proposed method, such as the cases in the second column of classes 2-6.

In general, three factors may influence the performance of our EFFNet, i.e., the dataset size that decides the network parameters, the motion blur noise that occurs during the collection of defect images from the production line, and the intricate back-

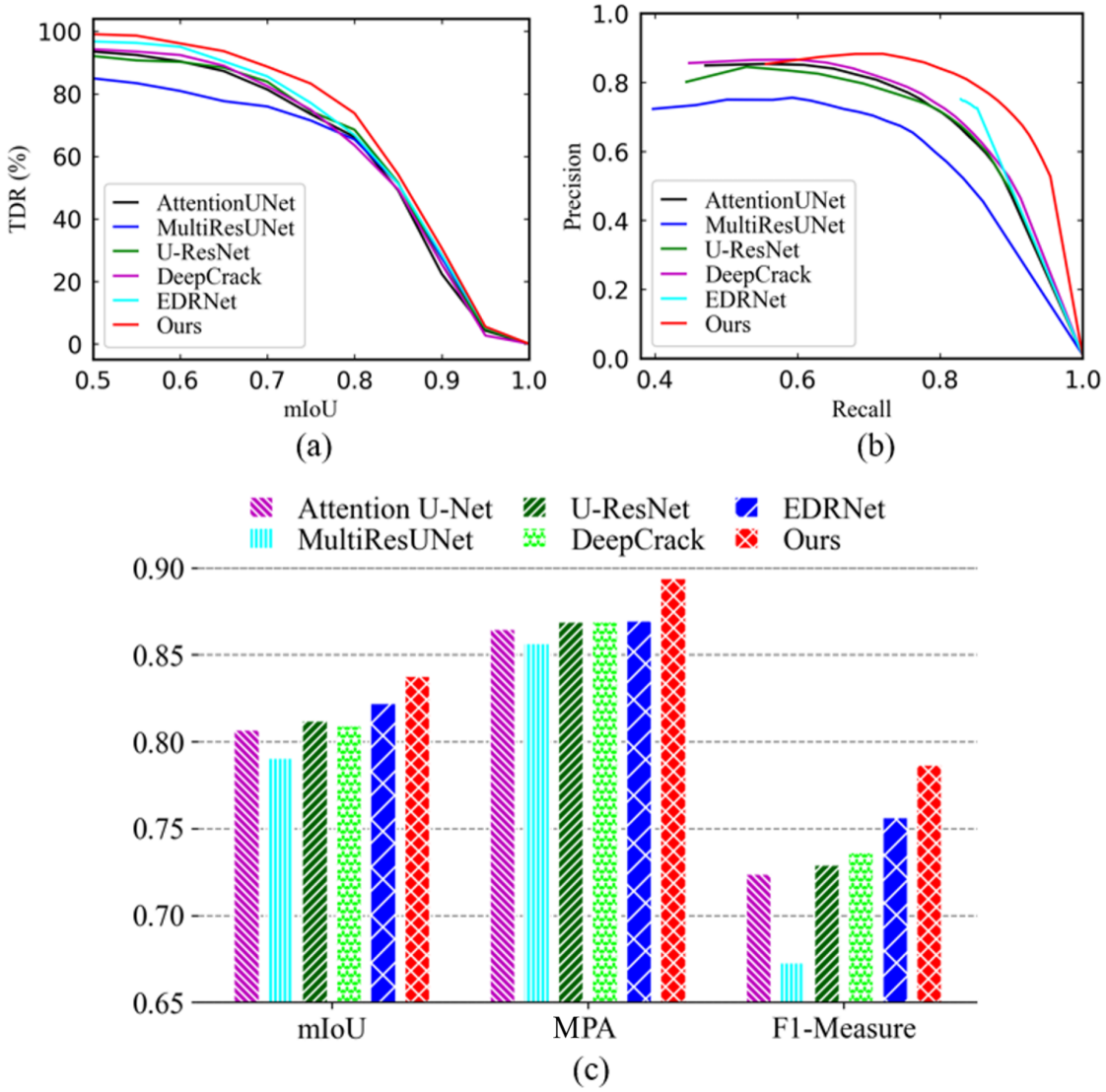

Figure 5.7: Evaluation results of our method and other CNNs. (a) TDR-mIoU curve. (b) P-R curve. (c) Histogram of the evaluation metrics.

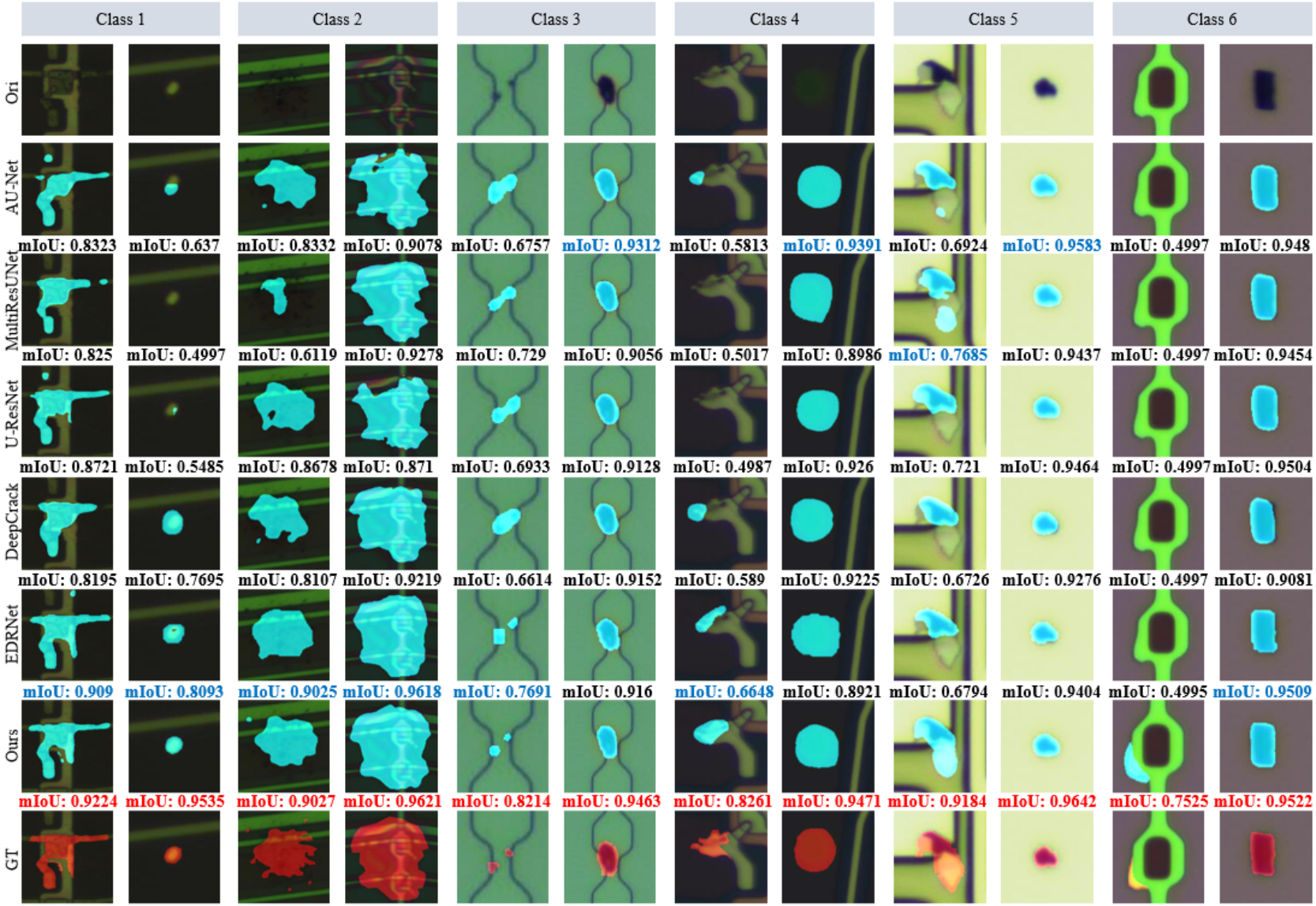

Figure 5.8: The visual comparison of our method with other state-of-the-art segmentation CNNs. Ori means original image, and GT represents the ground truth. AU-Net is Attention U-Net. The numbers underneath the images are the corresponding mIoU values of the detection, where the values in red are the best, and the ones in blue are the second best.

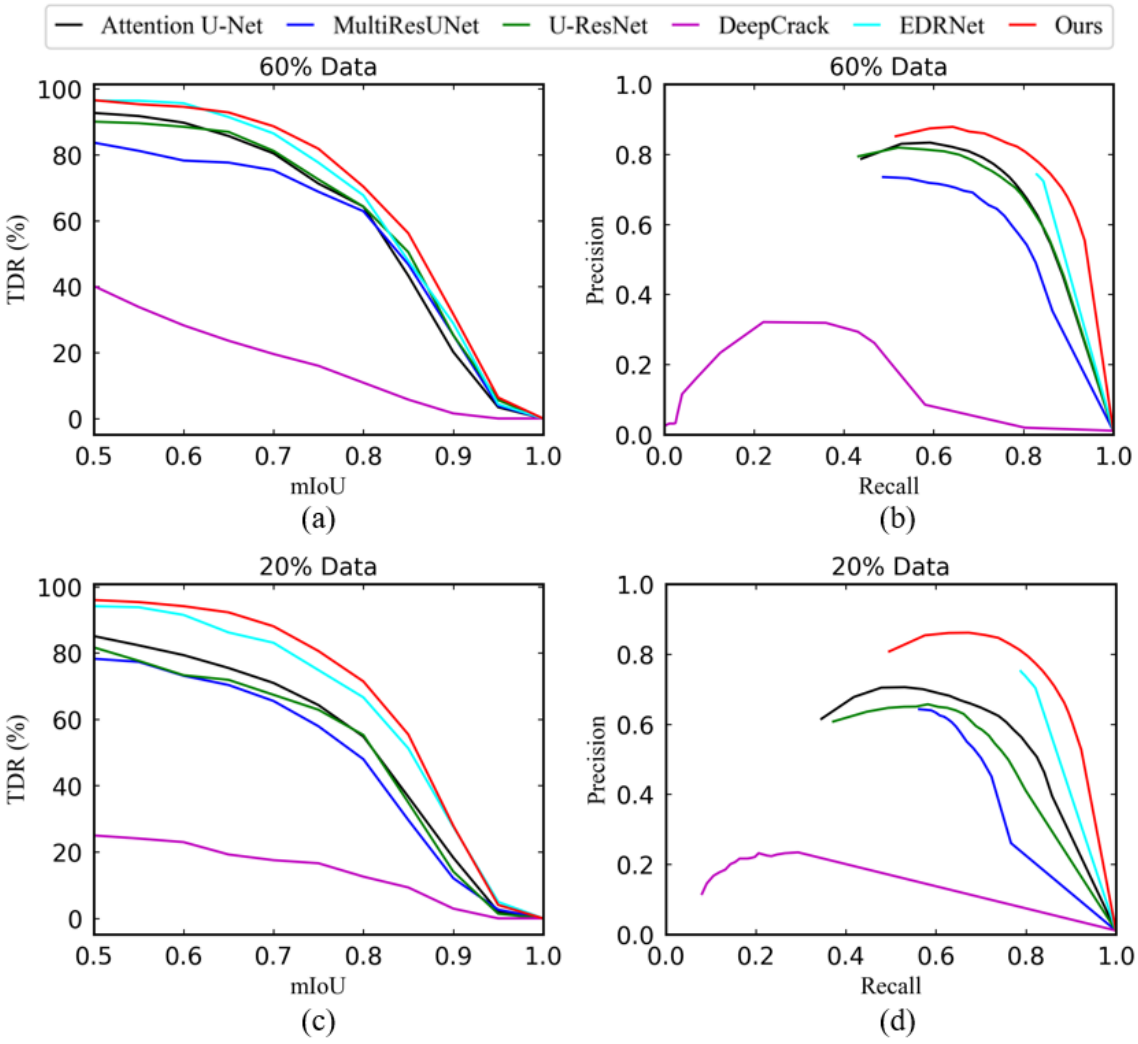

Figure 5.9: Evaluation curves of CNN-based methods trained with two sizes of datasets. (a) and (c) TDR-mIoU curve. (b) and (d) P-R curve.

grounds around the defects. We implemented the following experiments to investigate these factors and verify the robustness of our network.

Effects of Dataset Size

We set two groups of sub-training sets randomly extracted from the 100% training dataset at percentages of 20% and 60%, respectively, to train all the CNN-based methods. The evaluation results of the detection are given in Table 5.6. We can see that our network maintains the best performance against all the competitive methods in terms of the three metrics in both 60% and 20% dataset cases. It is fairly in conformity with the TDR-mIoU and P-R curves depicted in Fig. 5.9, where we can also see that the DeepCrack is quite sensitive to dataset size. Specifically, the proposed EFFNet has an improvement of at least 1.36% (in 60% dataset case), 2.3% (in 20% dataset case), and 2.04% (in 60% dataset case) on mIoU, MPA, and F1-Measure, respectively, compared to the second-best model EDRNet. The interesting part, though, is that the MPA of our EFFNet model has an enhancement of 0.1% in the 60% dataset case compared to the 100% one. We conjecture the cause is that the positive and negative samples in the defective images are highly imbalanced, while the latter occupies the majority. Moreover, although 100% of the data set enables the network to find more defects, it also increases the false-positive rate to a certain extent, which reduces the MPA.

Table 5.6: Evaluation results of deep-learning methods in the study of dataset size.

	60%			20%		
Metrics	mIoU	MPA	F1-Measure	mIoU	MPA	F1-Measure
Attention U-Net	0.7992	0.8569	0.7099	0.7647	0.8233	0.6323
MultiResUNet	0.7807	0.8457	0.6539	0.7387	0.7869	0.5773
U-ResNet	0.8052	0.8599	0.7141	0.7499	0.8139	0.5981
DeepCrack	0.5779	0.6594	0.2142	0.566	0.6042	0.1725
EDRNet	0.8214	0.867	0.7559	0.8136	0.8638	0.7355
Ours	**0.8350**	**0.8951**	**0.7763**	**0.8301**	**0.8868**	**0.7691**

Robustness to Motion Blur Noise

The defective images of display panels gathered from the production line are more or less subject to motion blur noise due to the relative movement between images and the camera. In order to study the robustness of our EFFNet model to this noise, we add two different degrees (Low and High) of motion blur to the input images. Fig. 5.10 presents the evaluation curves of the CNN-based methods in these two groups of experiments. It is observed from Fig. 5.10(a) and Fig. 5.10(b) that our approach outperforms all the other methods in terms of the TDR-mIoU and P-R curves when the motion blur noise is low. For the input images with high motion blur noise, although the TDR-mIoU curve of EDRNet is slightly higher than that of the EFFNet, its precision is smaller over a large range of thresholds, as shown in Fig. 5.10(c) and Fig. 5.10(d). The quantitative evaluation metrics of these methods are detailed in Table 5.7. Similarly, our model has the best performance in the group of low motion blur, 1.15%, 1.44%,

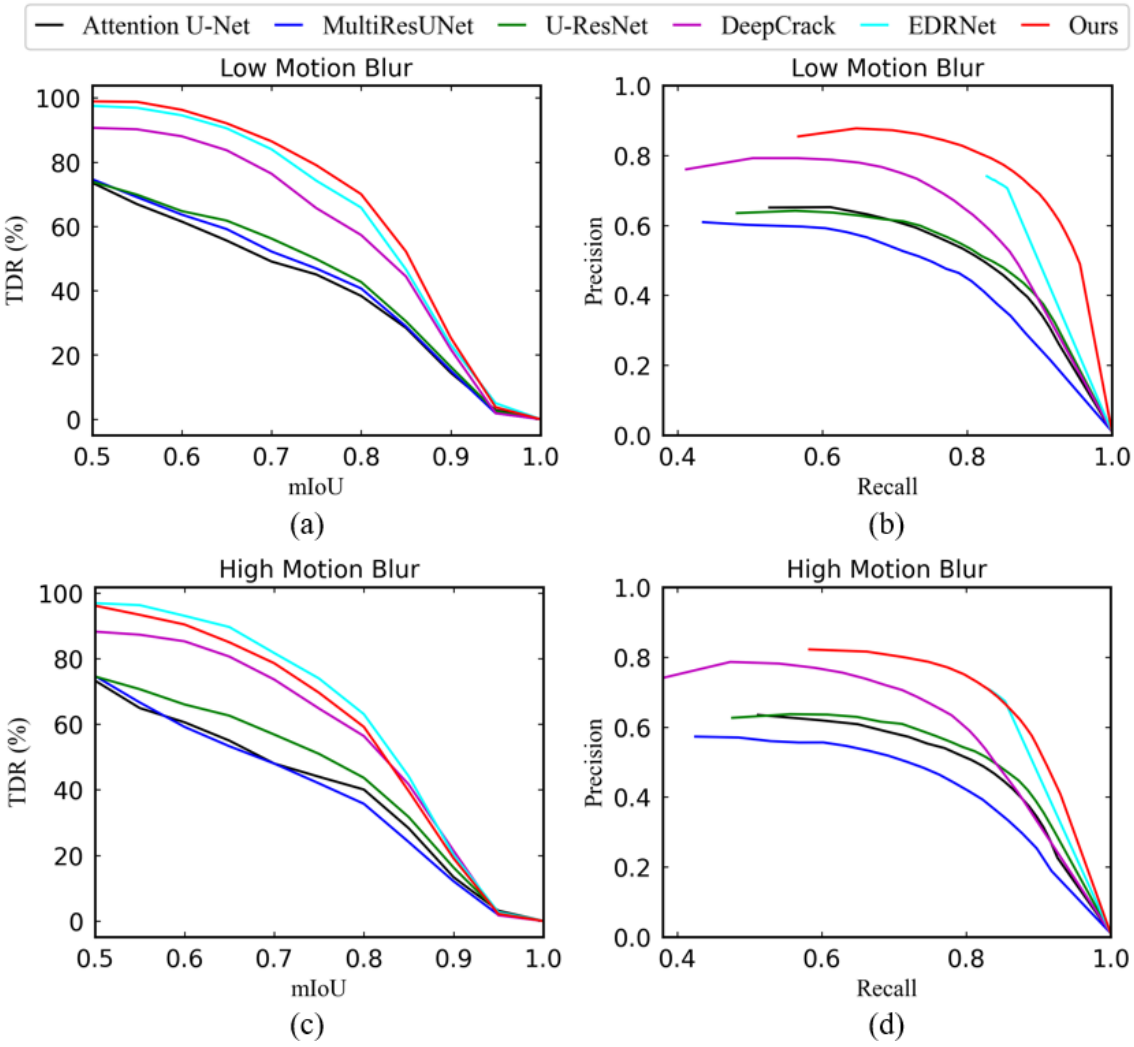

Figure 5.10: Evaluation curves of CNN-based methods in the robustness experiment. (a) and (c) TDR-mIoU curve. (b) and (d) P-R curve.

and 2.17% higher than the second-best model EDRNet in mIoU, MPA, and F1-Measure, respectively. However, Our model becomes the second-best method in the group of high motion blur, and EDRNet is the best one instead. It means that, compared to EDRNet, the highest efficiency of our EFFNet model (159 fps/s) is at the cost of its lower robustness against the high motion blur noise, while the efficiency of the former is much lower (60 fps/s).

Table 5.7: Evaluation results of deep-learning methods in the study of motion blur noise.

	Low motion blur			High motion blur		
Metrics	mIoU	MPA	F1-Measure	mIoU	MPA	F1-Measure
Attention U-Net	0.6694	0.7473	0.5	0.6646	0.7434	0.4895
MultiResUNet	0.7023	0.7561	0.5136	0.6825	0.7419	0.4806
U-ResNet	0.6756	0.7717	0.5346	0.6745	0.7747	0.53
DeepCrack	0.7861	0.8428	0.689	0.7779	0.831	0.6697
EDRNet	0.8161	0.8624	0.7488	**0.8067**	**0.8475**	**0.7335**
Ours	**0.8276**	**0.8768**	**0.7705**	0.7943	0.8318	0.7079

Performance in Different Backgrounds

The images used in our experiments have six classes of backgrounds containing different colors and textures (see Chapter 2). To study the effects of these different complex backgrounds on the defect detection performance of the deep-learning methods, we conduct the detection experiment of six sub-datasets, each of which contains images of the same class of background. The TDR-mIoU curves of the CNN-based approaches in this experiment are plotted in Fig. 5.11. Obviously, our EFFNet model performs the best almost over the whole mIoU threshold range in class 1 and class 6. For the other four classes, our model only outperforms other state-of-the-art segmentation methods in certain mIoU thresholds. Fig. 5.12 gives the P-R curves of the experiment, from which we can see that the proposed method has the best performance in class 1, class 2, class 4, and class 6. For class 3 and class 5, the proposed EFFNet has the highest precision over a certain range.

In addition, Table 5.8 summarizes the detail of the evaluation metrics of these methods in this experiment. We can observe that our model achieves the best performance in class 1, class 2, class 4, and class 6, which conforms with the analysis of the P-R curves mentioned above. Yet, EDRNet outperforms other methods in class 3 and class 5, and our method turns out to be the second-best. In particular, compared to the performance in other classes, the proposed model has the biggest mIoU and F1-Measure in class 2, yet the biggest MPA in class 3, i.e., 0.8709, 0.8454, and 0.9313, respectively. On the other hand, our model has the worst performance in class 3 in terms of mIoU and F1-Measure, yet the worst MPA in class 1 compared to its performance in other classes. The values are 0.8048, 0.7214, and 0.8282, respectively. It is worth noting that the inconsistency of MPA in class 1 and class 3 is also caused by the imbalance classification and the increase in the false-positive rate.

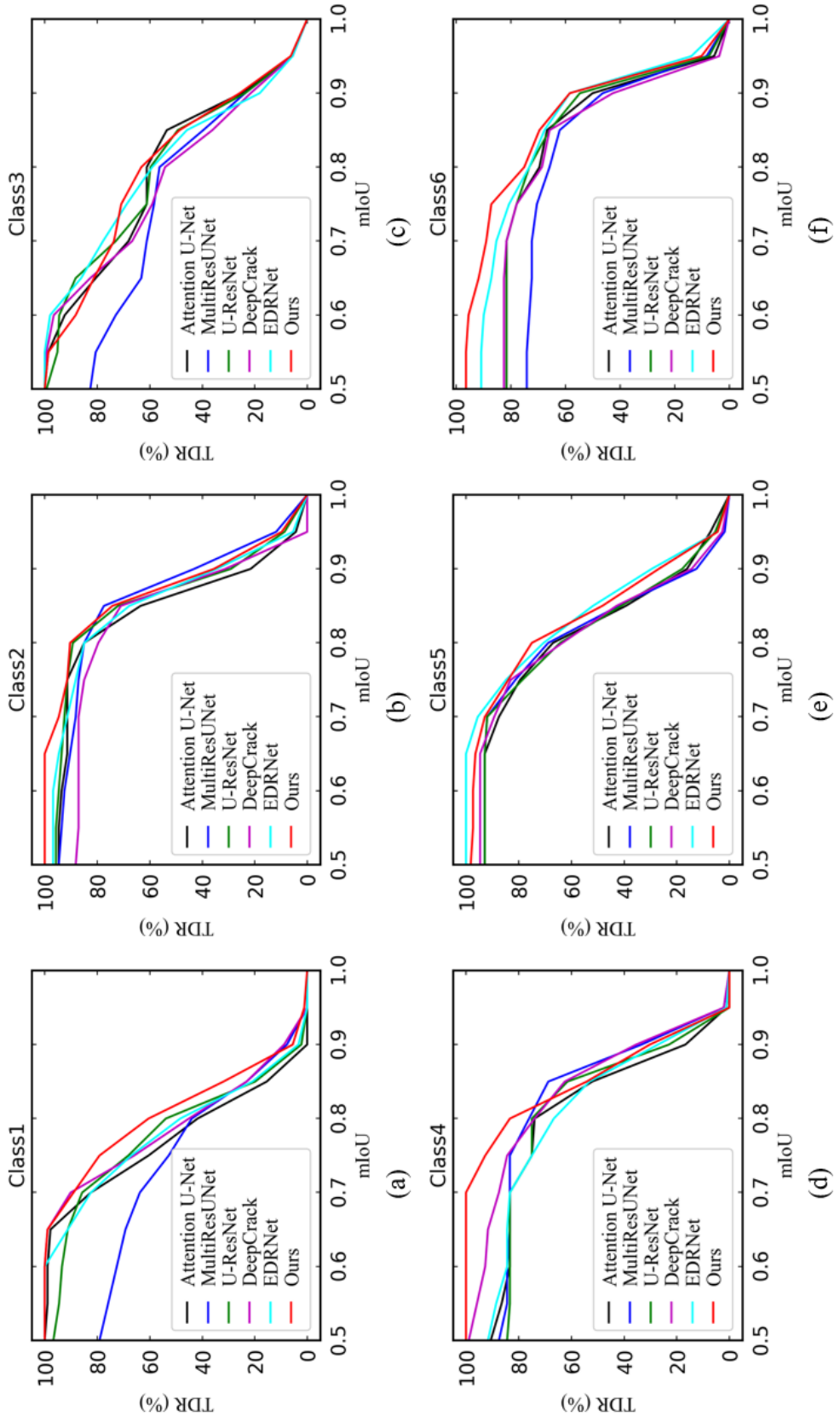

Figure 5.11: TDR-mIoU curves of CNN-based methods in the study of different backgrounds.

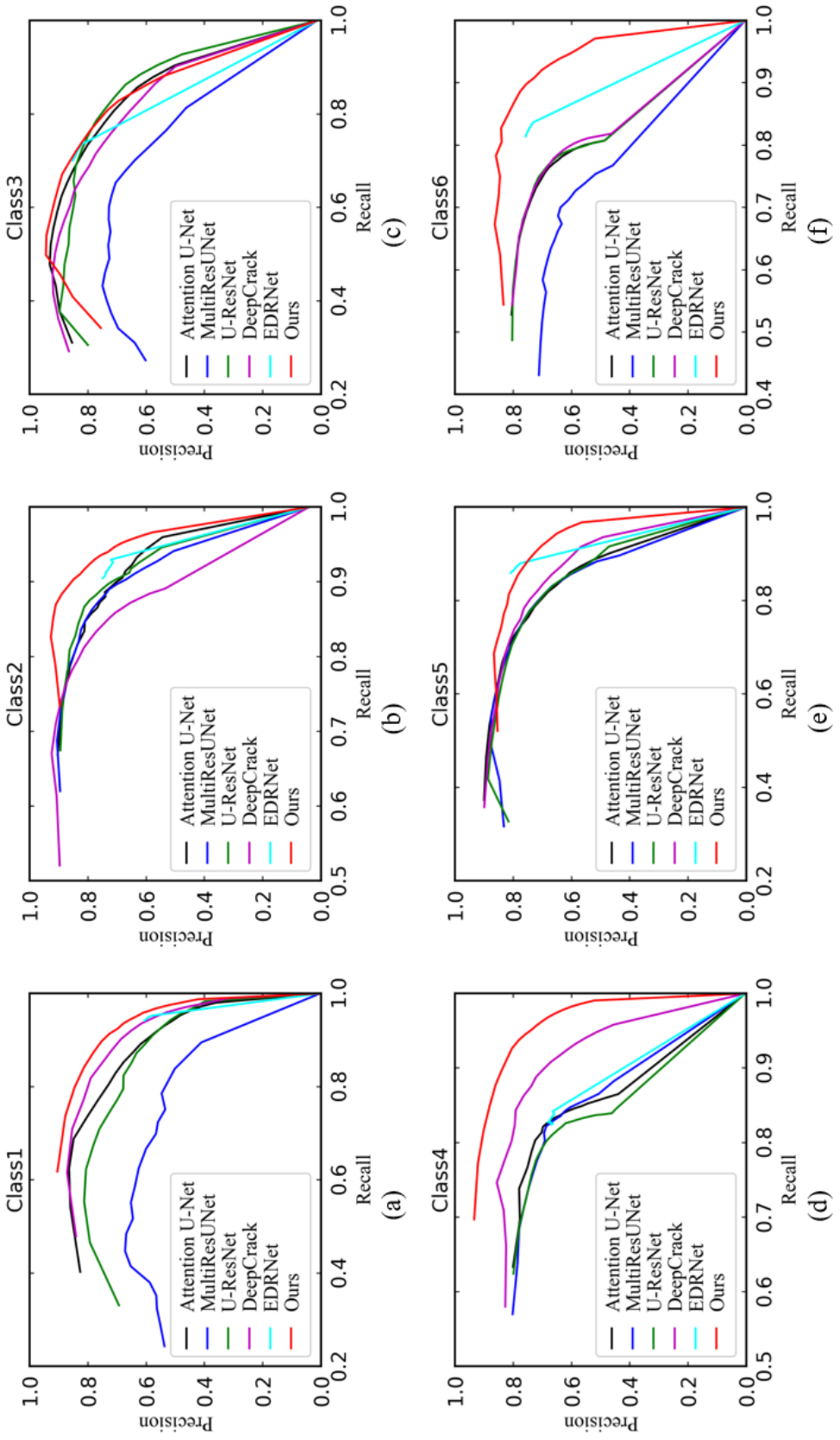

Figure 5.12: P-R curves of CNN-based methods in the study of different backgrounds.

Table 5.8: Evaluation results of deep-learning methods in the study of different backgrounds.

	Class1			Class2			Class3		
Metrics	mIoU	MPA	F1-Measure	mIoU	MPA	F1-Measure	mIoU	MPA	F1-Measure
Attention U-Net	0.7776	0.8229	0.7043	0.8398	0.8686	0.7872	0.7958	0.9238	0.7071
MultiResUNet	0.7283	0.831	0.5619	0.8554	0.8917	0.8014	0.7501	0.8598	0.5931
U-ResNet	0.7845	**0.8395**	0.7042	0.8524	0.8849	0.8066	0.7974	0.9196	0.712
DeepCrack	0.7903	0.8092	0.726	0.8231	0.8818	0.78	0.7852	0.91	0.6981
EDRNet	0.7866	0.8031	0.714	0.8506	0.8664	0.8084	**0.8049**	0.918	**0.7339**
Ours	**0.8108**	0.8282	**0.7567**	**0.8709**	**0.9043**	**0.8454**	0.8048	**0.9313**	0.7214
	Class4			Class5			Class6		
Metrics	mIoU	MPA	F1-Measure	mIoU	MPA	F1-Measure	mIoU	MPA	F1-Measure
Attention U-Net	0.8015	0.8286	0.7047	0.8044	0.8657	0.7243	0.8175	0.8389	0.7092
MultiResUNet	0.8224	0.8466	0.7284	0.8087	0.8787	0.7322	0.7906	0.8286	0.6444
U-ResNet	0.8071	0.8144	0.7039	0.8091	0.8773	0.7314	0.8243	0.8478	0.7184
DeepCrack	0.8366	0.8687	0.7792	0.8144	0.8869	0.7481	0.814	0.8369	0.7079
EDRNet	0.8049	0.8362	0.711	**0.8469**	**0.8964**	**0.8092**	0.8465	0.8749	0.7741
Ours	**0.8561**	**0.8794**	**0.827**	0.8307	0.8934	0.775	**0.861**	**0.8986**	**0.8115**

5.4.4 Analysis of Failure Cases

As analyzed in previous sections, the proposed network outperforms other state-of-the-art methods. Nevertheless, there are some cases still challenging for our network and the competitive methods. Fig. 5.13 shows some examples of the failed segmentation results of the proposed EFFNet, which to some extent reveal certain drawbacks of our model.

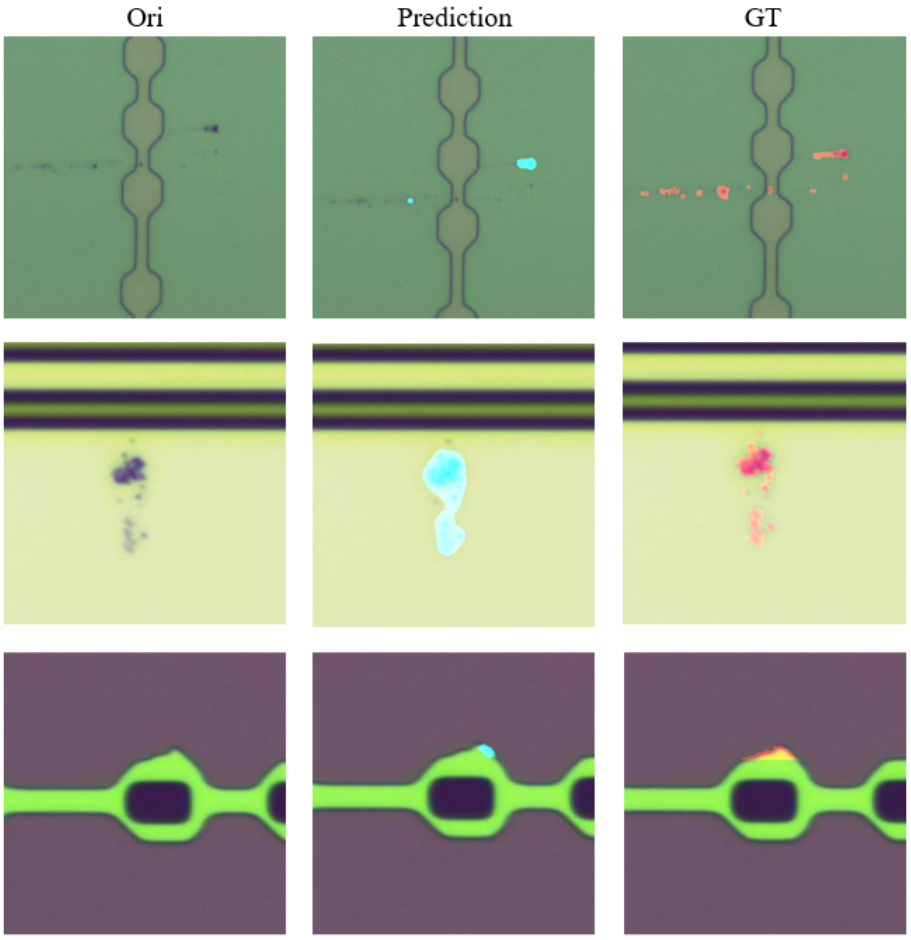

Figure 5.13: Failures of the proposed method.

First, the network misses some regions of interest where the contrast is mighty low, as shown in the first row of Fig. 5.13. Although these areas can be detected based on a lower threshold (fixed as 0.5 in our task) for the positive pixels, it will bring more false positives and reduce the accuracy. Furthermore, the network mistakenly treats some areas next to the target regions as defects, such as the example in the second row. We infer that this phenomenon resulted from lacking a large training dataset with enough numbers and variety, which prevents the network from learning more information related to the edge of defects. Finally, as can be observed in the last row of Fig. 5.13, the model fails to detect the malformation defect having a tiny size. The conjecture is that, in addition to the lack of a large training dataset, most of the information related to this defect is lost during the down-sampling process due to its small size. To address these problems, we have to obtain a dataset with enough diversity and amount at first,

which is a task for us in the future.

5.5 Conclusion

We proposed a novel EFFNet model for online defect detection that is of great importance for quality control and yield rate improvement of display panels. The modified ImageNet-trained VGG-16 ConvBlocks and a fine-tune strategy were introduced for the encoder to extensively extract intricate defect features. Furthermore, an element-wise feature fusion module (i.e., EFFM) based on the additive attention mechanism was developed for our decoder to fuse multi-level features to enhance the detection accuracy while avoiding more computational complexity. Experimental results show that our method performs better than the state-of-the-art defect detection methods and other segmentation techniques. It also has good robustness against motion blur noise and small training dataset. Moreover, our model can detect defects at speeds acceptable for real-time defect detection.

Nevertheless, there are still some improvements to be achieved to put the developed network into practice. In future work, it would be interesting to design a more efficient data enhancement strategy to avoid the failure of the proposed method in more challenging cases and optimize the framework to achieve even higher accuracy and faster speed for the detection of different defects in display panels.

Chapter 6

General Conclusions and Perspectives

Conclusions

In this thesis, we described an original work of developing effective deep learning methods for explainable classification of histopathological breast cancer images and display panel online defect detection. To better and more comprehensively present our work, we first gave a brief introduction about the problem in Chapter 1, which leads to our objectives. Furthermore, the objects to be investigated and the related background were introduced in Chapter 2, including breast cancer diagnosis, display panel defect detection, and relevant state-of-the-art deep learning methods. The main contributions presented in Chapter 3 to Chapter 5 are listed as follows:

- In Chapter 3, we proposed a configurable convolutional neural network (ConfigNet) capable of transforming into different configurations adapted to different tasks and objects. The ConfigNet mainly contains two functional configurations: 1) feature extraction module (FEM)-decision map generator (DMG)-classifier configuration for explainable image classification, 2) encoder-decoder configuration for object segmentation and localization. The backbones of the FEM and encoder of these two configurations are constructed through the transfer learning of existing CNNs with deep convolutional layers. Extensive experiments of the FEM-DMG-classifier configuration (MICNet) based on multiple instance learning (MIL), VGG11, a basic DMG, and a weighted average pooling (WAP) classifier on the BreakHis and Camelyon16 patch-based datasets demonstrated that our MICNet outperforms other CNN models in the classification of histopathological breast cancer images and can provide a logical visual explanation that supports the network prediction. Moreover, the experimental results of the encoder-decoder configuration (SCAFFNet) based on spatial and channel attention-guided feature fusion module (SCAFFM) and decoder with a "bottleneck" structure on the Camelyon16 patch-based dataset indicated that the SCAFFNet outperforms the state-of-the-art segmentation models in breast tumor segmentation, especially in the challenging case where the breast tumors have complex boundaries.

- In Chapter 4, we proposed a novel weakly-supervised learning-based network named ExplaCNet that is built on the FEM-DMG-classifier configuration of our

ConfigNet to achieve the explainable classification of histopathological breast cancer images. The MIL that encourages the network to identify more normal tissues and the WAP classifier that forces the ExplaCNet to learn to recognize more lesion tissues were adopted. In particular, we developed a DMG with multi-scale convolution filters that allow a reasonable compromise in the ability of ExplaCNet to identify normal and lesion tissues to generate refined decision maps for final classification and visual explanation. Experimental results on Camelyon16 patch-based dataset and BreakHis dataset showed that our ExplaCNet outperforms state-of-the-art explainable methods in terms of visual explanation while keeping a competitive classification performance.

- In Chapter 5, we proposed a novel element-wise feature fusion network (EFFNet) based on the encoder-decoder configuration of the ConfigNet to achieve high-accuracy real-time defect detection of display panels. The method adopted a transfer learning and fine-tuning strategy for feature extraction layers and a decoder with relatively less computational complexity. Particularly, a feature fusion module based on element-wise addition of pyramid features was proposed in skip connection to improve detection efficiency and accuracy. The EFFNet was compared with traditional defect detection techniques and state-of-the-art CNN-based models. Extensive experiments demonstrated that the EFFNet can accurately detect defects with complex textures, ambiguous boundaries and low contrast. It also has good robustness against motion blur noise. It outperforms state-of-the-art methods in terms of mIoU, MPA, and F1-Measure. Moreover, it is able to detect defects at speeds of up to 159 fps/s with input images of size 256×256.

In summary, we proposed a configurable convolutional neural network named ConfigNet that can be configured into different configurations suitable for multiple tasks and objects, such as the explainable classification of histopathological breast cancer images and the online defect detection of display panels. The proposed methods (configurations) present superior performances in breast cancer images explainable classification that has considerable potential for supporting the deployment of deep learning-based CAD systems in real-world clinical settings as well as in display panel online defect detection that is of great importance for improving the yield rate.

Perspectives

Considering the limitations found in the current study, future work would be in two folds. The first one is breast cancer diagnosis involving the investigation of images of other modalities apart from histopathological images, the classification of sub-categories of breast cancer, and the improvement of the FEM-DMG-classifier configuration. The other is display panel online defect detection involving the investigation of more real-world display panel defect samples and the improvement of the encoder-decoder configuration.

For breast cancer diagnosis, future work may include:

- Investigating the explainable classification of breast cancer images of other modalities, including mammogram, ultrasonography, CT, MRI, and PET-CT based on the FEM-DMG-classifier configuration of our ConfigNet to comprehensively study the role and contribution of ConfigNet to the detection and diagnosis of breast cancer, which further facilitates the deployment of deep learning-based CAD systems in clinical settings.

- Pushing the study of the explainable classification of breast cancer deeper into sub-categories to complement the capability of ConfigNet in the diagnosis of all kinds of breast cancer categories, thus more comprehensively assisting clinical experts in making optimal treatment proposals. It requires more detailed annotations, i.e., image-level labels of the sub-categories and their pixel-level masks, that need plenty of workloads.

- Putting the developed ExplaCNet into practice and investigating the optimal trade-off between generating better human-understandable explanation maps and obtaining more accurate classification results in the clinical setting would help to fulfill its clinical application. Furthermore, the idea of the Transformer (Vaswani et al., 2017) might be helpful for extending the FEM-DMG-classifier configuration of the ConfigNet to achieve more breast cancer analysis tasks.

For display panel online defect detection, future work could be:

- Obtain datasets with more diversity and complexity gathered from the factory production line and investigate the performance of the encoder-decoder configuration of our ConfigNet on these datasets. It would help improve its performance and enhance its reliability and stability. Furthermore, more studies on the backbone of the encoder-decoder configuration based on lighter models such as SqueezeNet (Iandola et al., 2016) might be helpful for its integration into AOI systems of terminal usage.

List of Publications

Journal Papers:

- **He, F.**, Tan, J., Wang, W., Liu, S., Zhu, Y. and Liu, Z., 2022. EFFNet: Element-wise feature fusion network for defect detection of display panels. Signal Processing: Image Communication, Under Review.
- **He, F.**, Wang, W., Nanding, A., Kuai, Z., Li, X., Zhu, Y. and Liu, Z., 2022. Explainable classification of histopathological breast cancer images based on weakly-supervised learning. Medical Image Analysis, Under Review.
- **He, F.**, Wang, W., Zhu, Y. and Liu, Z., 2023. SCAFFNet: Spatial and channel attention-guided feature fusion network for medical image segmentation. Journal of Biomedical Informatics, Waiting to be submitted.

Conference Paper:

- **He, F.**, Zhu, Y., Wang, W., Nanding, A., Kuai, Z., Li, X. and Liu, Z., 2022. Multi-Instance Classification of Histopathological Breast Cancer Images with Visual Explanation. In 2022 16th IEEE International Conference on Signal Processing (ICSP), pp. 431-436.

Bibliography

Adoui, M. E., Drisis, S., and Benjelloun, M. (2020). Multi-input deep learning architecture for predicting breast tumor response to chemotherapy using quantitative MR images. *International Journal of Computer Assisted Radiology and Surgery*, 15(9):1491–1500. 105

Akram, M., Iqbal, M., Daniyal, M., and Khan, A. U. (2017). Awareness and current knowledge of breast cancer. *Biological research*, 50(1):1–23. 1, 33, 41

Alom, M. Z., Hasan, M., Yakopcic, C., Taha, T. M., and Asari, V. K. (2018). Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. *arXiv preprint arXiv:1802.06955*. 4, 16, 35, 67, 78, 98, 99

AlZubaidi, A. K., Sideseq, F. B., Faeq, A., and Basil, M. (2017). Computer aided diagnosis in digital pathology application: Review and perspective approach in lung cancer classification. In *2017 annual conference on new trends in information & Communications technology applications (NTICT)*, pages 219–224. IEEE. 104

Amedee, R. G. and Dhurandhar, N. R. (2001). Fine-needle aspiration biopsy. *The Laryngoscope*, 111(9):1551–1557. 51

Antoch, G., Stattaus, J., Nemat, A. T., Marnitz, S., Beyer, T., Kuehl, H., Bockisch, A., Debatin, J. F., and Freudenberg, L. S. (2003). Non–small cell lung cancer: dual-modality pet/ct in preoperative staging. *Radiology*, 229(2):526–533. 50

Apesteguía, L. and Pina, L. J. (2011). Ultrasound-guided core-needle biopsy of breast lesions. *Insights into imaging*, 2(4):493–500. 51

Ashual, O. and Wolf, L. (2019). Specifying object attributes and relations in interactive scene generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4561–4569. 104

Aswathy, M. and Jagannath, M. (2017). Detection of breast cancer on digital histopathology images: Present status and future possibilities. *Informatics in Medicine Unlocked*, 8:74–79. 104

Avril, N., Dose, J., Jänicke, F., Bense, S., Ziegler, S., Laubenbacher, C., Römer, W., Pache, H., Herz, M., Allgayer, B., et al. (1996). Metabolic characterization of breast tumors with positron emission tomography using f-18 fluorodeoxyglucose. *Journal of clinical oncology*, 14(6):1848–1857. 50

Avril, N., Schelling, M., Dose, J., Weber, W. A., and Schwaiger, M. (1999). Utility of pet in breast cancer. *Clinical Positron Imaging*, 2(5):261–271. 50

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(12):2481–2495. xii, 4, 16, 36, 66, 67, 78, 98, 99, 104, 127

Baek, S.-I., Kim, W.-S., Koo, T.-M., Choi, I., and Park, K.-H. (2004). Inspection of defect on lcd panel using polynomial approximation. In *2004 IEEE Region 10 Conference TENCON 2004.*, pages 235–238. IEEE. 58

Barnes, B. M., Sohn, M. Y., Goasmat, F., Zhou, H., Vladár, A. E., Silver, R. M., and Arceo, A. (2013). Three-dimensional deep sub-wavelength defect detection using λ= 193 nm optical microscopy. *Opt. Express*, 21(22):26219–26226. 126

Bayramoglu, N., Kannala, J., and Heikkilä, J. (2016a). Deep learning for magnification independent breast cancer histopathology image classification. In *2016 23rd International conference on pattern recognition (ICPR)*, pages 2440–2445. IEEE. 2, 34, 89, 90

Bayramoglu, N., Kannala, J., and Heikkila, J. (2016b). Deep learning for magnification independent breast cancer histopathology image classification. In *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE. 118

Bazira, P. J., Ellis, H., and Mahadevan, V. (2021). Anatomy and physiology of the breast. *Surgery (Oxford)*. xi, 8, 40

Bejnordi, B. E., Veta, M., van Diest, P. J., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J. A. W. M., Hermsen, M., Manson, Q. F., Balkenhol, M., Geessink, O., Stathonikos, N., van Dijk, M. C., Bult, P., Beca, F., Beck, A. H., Wang, D., Khosla, A., Gargeya, R., Irshad, H., Zhong, A., Dou, Q., Li, Q., Chen, H., Lin, H.-J., Heng, P.-A., Haß, C., Bruni, E., Wong, Q., Halici, U., Ümit Öner, M., Cetin-Atalay, R., Berseth, M., Khvatkov, V., Vylegzhanin, A., Kraus, O., Shaban, M., Rajpoot, N., Awan, R., Sirinukunwattana, K., Qaiser, T., Tsang, Y.-W., Tellez, D., Annuscheit, J., Hufnagl, P., Valkonen, M., Kartasalo, K., Latonen, L., Ruusuvuori, P., Liimatainen, K., Albarqouni, S., Mungal, B., George, A., Demirci, S., Navab, N., Watanabe, S., Seno, S., Takenaka, Y., Matsuda, H., Phoulady, H. A., Kovalev, V., Kalinovsky, A., Liauchuk, V., Bueno, G., Fernandez-Carrobles, M. M., Serrano, I., Deniz, O., Racoceanu, D., and and, R. V. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199. 9, 52

Belharbi, S., Rony, J., Dolz, J., Ayed, I. B., McCaffrey, L., and Granger, E. (2021). Deep interpretable classification and weakly-supervised segmentation of histology images via max-min uncertainty. *IEEE Transactions on Medical Imaging*, 41(3):702–714. 105, 110

Berg, W. A., Blume, J. D., Cormack, J. B., Mendelson, E. B., Lehrer, D., Böhm-Vélez, M., Pisano, E. D., Jong, R. A., Evans, W. P., Morton, M. J., et al. (2008). Combined screening with ultrasound and mammography vs mammography alone in women at elevated risk of breast cancer. *Jama*, 299(18):2151–2163. 47

Bissi, L., Baruffa, G., Placidi, P., Ricci, E., Scorzoni, A., and Valigi, P. (2013). Automated defect detection in uniform and structured fabrics using gabor filters and pca. *Journal of Visual Communication and Image Representation*, 24(7):838–845. 58

Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*. 16, 78

Boone, J. M., Kwan, A. L., Yang, K., Burkett, G. W., Lindfors, K. K., and Nelson, T. R. (2006). Computed tomography for imaging the breast. *Journal of mammary gland biology and neoplasia*, 11(2):103–111. 2, 34

Bottou, L. (2012). Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, pages 421–436. Springer. 69

Brodersen, J. and Siersma, V. D. (2013). Long-term psychosocial consequences of false-positive screening mammography. *The Annals of Family Medicine*, 11(2):106–115. 46

Buck, A., Schirrmeister, H., Kühn, T., Shen, C., Kalker, T., Kotzerke, J., Dankerl, A., Glatting, G., Reske, S., and Mattfeldt, T. (2002). Fdg uptake in breast cancer: correlation with biological and clinical prognostic parameters. *European journal of nuclear medicine and molecular imaging*, 29(10):1317–1323. 50

Burstein, H. J., Polyak, K., Wong, J. S., Lester, S. C., and Kaelin, C. M. (2004). Ductal carcinoma in situ of the breast. *New England Journal of Medicine*, 350(14):1430–1441. 44

Carbonneau, P. E., Dugdale, S. J., Breckon, T. P., Dietrich, J. T., Fonstad, M. A., Miyamoto, H., and Woodget, A. S. (2020). Adopting deep learning methods for airborne rgb fluvial scene classification. *Remote Sens. Environ.*, 251:112107. 126

Carneiro, G., Nascimento, J., and Bradley, A. P. (2017). Automated analysis of unregistered multi-view mammograms with deep learning. *IEEE transactions on medical imaging*, 36(11):2355–2365. 16, 78

Chan, H.-P., Hadjiiski, L. M., and Samala, R. K. (2020). Computer-aided diagnosis in the era of deep learning. *Medical Physics*, 47(5). 104

Chang, Y.-C., Chang, K.-H., Meng, H.-M., and Chiu, H.-C. (2022). A novel multicategory defect detection method based on the convolutional neural network method for tft-lcd panels. *Math. Probl. Eng.*, 2022. 4, 35, 127

Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. (2018). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE. xii, 73, 105, 110, 113, 114

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848. 2, 34

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2018). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848. 104

Chikontwe, P., Kim, M., Nam, S. J., Go, H., and Park, S. H. (2020). Multiple instance learning with center embeddings for histopathology classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 519–528. Springer. 84

Choi, D.-C., Jeon, Y.-J., Lee, S. J., Yun, J. P., and Kim, S. W. (2014). Algorithm for detecting seam cracks in steel plates using a gabor filter combination method. *Appl. Opt.*, 53(22):4865–4872. 126

Chu, M., Gong, R., Gao, S., and Zhao, J. (2017). Steel surface defects recognition based on multi-type statistical features and enhanced twin support vector machine. *Chemomet. Intell. Lab. Syst.*, 171:140–150. 4, 35, 126

Chuba, P. J., Hamre, M. R., Yap, J., Severson, R. K., Lucas, D., Shamsa, F., and Aref, A. (2005). Bilateral risk for subsequent breast cancer after lobular carcinoma-in-situ: analysis of surveillance, epidemiology, and end results data. *Journal of Clinical Oncology*, 23(24):5534–5541. 44

Ciga, O. and Martel, A. L. (2021). Learning to segment images with classification labels. *Medical Image Analysis*, 68:101912. xii, 75, 105, 113, 114

Cong, L., Feng, W., Yao, Z., Zhou, X., and Xiao, W. (2020). Deep learning model as a new trend in computer-aided diagnosis of tumor pathology for lung cancer. *Journal of Cancer*, 11(12):3615. 2, 34

Cooper, H. S., Patchefsky, A. S., and Krall, R. A. (1978). Tubular carcinoma of the breast. *Cancer*, 42(5):2334–2342. 42

De Bresser, J., De Vos, B., Van der Ent, F., and Hulsewe, K. (2010). Breast mri in clinically and mammographically occult breast cancer presenting with an axillary metastasis: a systematic review. *European Journal of Surgical Oncology (EJSO)*, 36(2):114–119. 49

Delotte, J., Karimdjee, B. S., Cua, E., Pop, D., Bernard, J.-L., Bongain, A., and Benchimol, D. (2009). Gas gangrene of the breast: management of a potential life-threatening infection. *Archives of gynecology and obstetrics*, 279(1):79–81. 7, 41

desai, s. and Ramaswamy, H. G. (2020). Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 105, 110, 113, 114

Dhariwal, P. and Nichol, A. (2021). Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794. 104

Dietterich, T. G., Lathrop, R. H., and Lozano-Pérez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71. 83

Dika, E., Curti, N., Giampieri, E., Veronesi, G., Misciali, C., Ricci, C., Castellani, G., Patrizi, A., and Marcelli, E. (2022). Advantages of manual and automatic computer-aided compared to traditional histopathological diagnosis of melanoma: A pilot study. *Pathology - Research and Practice*, 237:154014. 104

Dixon, J., Anderson, T., Page, D., Lee, D., and Duffy, S. (1982). Infiltrating lobular carcinoma of the breast. *Histopathology*, 6(2):149–161. 41

Dong, H., Song, K., He, Y., Xu, J., Yan, Y., and Meng, Q. (2019). Pga-net: Pyramid feature fusion and global context attention network for automated surface defect detection. *IEEE Transactions on Industrial Informatics*, 16(12):7448–7458. 16, 78

Dong, Y., Liu, Q., Du, B., and Zhang, L. (2022). Weighted feature fusion of convolutional neural network and graph attention network for hyperspectral image classification. *IEEE Transactions on Image Processing*, 31:1559–1572. 2, 34

Dozat, T. (2016). Incorporating nesterov momentum into adam. 69

Durand, T., Mordan, T., Thome, N., and Cord, M. (2017). Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. xii, 2, 34, 72, 105, 113, 114

Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Liu, Y., Topol, E., Dean, J., and Socher, R. (2021). Deep learning-enabled medical computer vision. *NPJ digital medicine*, 4(1):1–9. 2, 34

Fahad Ullah, M. (2019). Breast cancer: current perspectives on the disease status. *Breast Cancer Metastasis and Drug Resistance*, pages 51–64. 7, 41

Falk, R. S., Hofvind, S., Skaane, P., and Haldorsen, T. (2013). Overdiagnosis among women attending a population-based mammography screening program. *International journal of cancer*, 133(3):705–712. 46

Ferguson, M., Ak, R., Lee, Y.-T. T., and Law, K. H. (2017). Automatic localization of casting defects with convolutional neural networks. In *2017 IEEE international conference on big data (big data)*, pages 1726–1735. IEEE. xi, xii, 13, 63

Fernández-Aguilar, S., Simon, P., Buxant, F., Simonart, T., and Noël, J.-C. (2005). Tubular carcinoma of the breast and associated intra-epithelial lesions: a comparative study with invasive low-grade ductal carcinomas. *Virchows Archiv*, 447(4):683–687. 42

Fukushima, K. and Miyake, S. (1982). Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer. 61

Gao, J., Li, D., and Havlin, S. (2014). From a single network to a network of networks. *National Science Review*, 1(3):346–356. 83

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (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. 16, 78

Gøtzsche, P. C. and Jørgensen, K. J. (2013). Screening for breast cancer with mammography. *Cochrane database of systematic reviews*, (6). 1, 33

Gour, M., Jain, S., and Kumar, T. S. (2020). Residual learning based CNN for breast cancer histopathological image classification. *International Journal of Imaging Systems and Technology*, 30(3):621–635. 2, 34, 104, 118

Gupta, G. P. and Massagué, J. (2006). Cancer metastasis: building a framework. *Cell*, 127(4):679–695. 9, 41

Han, G., Huang, S., Ma, J., He, Y., and Chang, S.-F. (2022). Meta faster r-cnn: Towards accurate few-shot object detection with attentive feature alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 780–789. 16, 78

Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *cell*, 100(1):57–70. 9, 41

Harris, L. N., Ismaila, N., McShane, L. M., Andre, F., Collyar, D. E., Gonzalez-Angulo, A. M., Hammond, E. H., Kuderer, N. M., Liu, M. C., Mennel, R. G., et al. (2016). Use of biomarkers to guide decisions on adjuvant systemic therapy for women with early-stage invasive breast cancer: American society of clinical oncology clinical practice guideline. *Journal of Clinical Oncology*, 34(10):1134. 1, 33

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969. 16, 78

He, K., Zhang, X., Ren, S., and Sun, J. (2016a). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778. xii, 2, 4, 16, 18, 34, 36, 64, 78, 79, 126

He, K., Zhang, X., Ren, S., and Sun, J. (2016b). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 71, 75, 104

He, Z. and Sun, L. (2015). Surface defect detection method for glass substrate using improved otsu segmentation. *Appl. Opt.*, 54(33):9823–9830. 3, 35, 126

Hecht-Nielsen, R. (1992). Theory of the backpropagation neural network. In *Neural networks for perception*, pages 65–93. Elsevier. 61

Heslin, M. J., Lewis, J. J., Woodruff, J. M., and Brennan, M. F. (1997). Core needle biopsy for diagnosis of extremity soft tissue sarcoma. *Annals of surgical oncology*, 4(5):425–431. 51

Heywang-Köbrunner, S. H., Hacker, A., and Sedlacek, S. (2011). Advantages and disadvantages of mammography screening. *Breast care*, 6(3):199–207. 46

Hoff, S. R., Abrahamsen, A.-L., Samset, J. H., Vigeland, E., Klepp, O., and Hofvind, S. (2012). Breast cancer: missed interval and screening-detected cancer at full-field digital mammography and screen-film mammography—results from a retrospective review. *Radiology*, 264(2):378–386. 46

Hofvind, S., Ponti, A., Patnick, J., Ascunce, N., Njor, S., Broeders, M., Giordano, L., Frigerio, A., and Törnberg, S. (2012). False-positive results in mammographic screening for breast cancer in europe: a literature review and survey of service screening programmes. *Journal of medical screening*, 19(1_suppl):57–66. 46

Hooley, R. J., Scoutt, L. M., and Philpotts, L. E. (2013). Breast ultrasonography: state of the art. *Radiology*, 268(3):642–659. 2, 34, 47

Hsu, W.-W., Wu, Y., Hao, C., Hou, Y.-L., Gao, X., Shao, Y., Zhang, X., He, T., and Tai, Y. (2021). A computer-aided diagnosis system for breast pathology: A deep learning approach with model interpretability from pathological perspective. 104

Hu, B. and Wang, J. (2020). Detection of pcb surface defects with improved faster-rcnn and feature pyramid network. *IEEE Access*, 8:108335–108345. 4, 35, 127

Hu, X., Villodre, E. S., Larson, R., Rahal, O. M., Wang, X., Gong, Y., Song, J., Krishnamurthy, S., Ueno, N. T., Tripathy, D., et al. (2021). Decorin-mediated suppression of tumorigenesis, invasion, and metastasis in inflammatory breast cancer. *Communications biology*, 4(1):1–14. 1, 33

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708. xii, 4, 16, 36, 64, 78

Huang, K., Misra, S., Bagaria, S. P., and Gabriel, E. M. (2021a). Outcomes of patients with invasive mucinous and tubular carcinomas of the breast. *The Breast Journal*, 27(9):691–699. 1, 33

Huang, M. L., Adrada, B. E., Candelaria, R., Thames, D., Dawson, D., and Yang, W. T. (2014). Stereotactic breast biopsy: pitfalls and pearls. *Techniques in vascular and interventional radiology*, 17(1):32–39. 51

Huang, Y. and Chung, A. C. S. (2019). Evidence localization for pathology images using weakly supervised learning. In *Lecture Notes in Computer Science*, pages 613–621. Springer International Publishing. xii, 75, 105, 110, 113, 114

Huang, Y., Jing, J., and Wang, Z. (2021b). Fabric defect segmentation method based on deep learning. *IEEE Transactions on Instrumentation and Measurement*, 70:1–15. 16, 78

Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106. 61

Hubel, D. H. and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243. 61

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., and Keutzer, K. (2016). Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. *arXiv preprint arXiv:1602.07360*. 31, 153

Ibtehaz, N. and Rahman, M. S. (2020). Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation. *Neural networks*, 121:74–87. 16, 67, 78, 138

Ilse, M., Tomczak, J., and Welling, M. (2018a). Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR. xii, 73, 74, 89

Ilse, M., Tomczak, J., and Welling, M. (2018b). Attention-based deep multiple instance learning. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2127–2136. PMLR. 105, 113, 114

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR. 13, 64

Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., and Maier-Hein, K. H. (2021). nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods*, 18(2):203–211. 126

Jacquemier, J., Padovani, L., Rabayrol, L., Lakhani, S. R., Penault-Llorca, F., Denoux, Y., Fiche, M., Figueiro, P., Maisongrosse, V., Ledoussal, V., et al. (2005). Typical medullary breast carcinomas have a basal/myoepithelial phenotype. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, 207(3):260–268. 42

Jafari, S. H., Saadatpour, Z., Salmaninejad, A., Momeni, F., Mokhtari, M., Nahand, J. S., Rahmati, M., Mirzaei, H., and Kianmehr, M. (2018). Breast cancer diagnosis: Imaging techniques and biochemical markers. *Journal of cellular physiology*, 233(7):5200–5213. 46

Jain, R. K. (2005). Normalization of tumor vasculature: an emerging concept in antiangiogenic therapy. *Science*, 307(5706):58–62. 9, 41

Jia, Z., Huang, X., Eric, I., Chang, C., and Xu, Y. (2017). Constrained deep weak supervision for histopathology image segmentation. *IEEE transactions on medical imaging*, 36(11):2376–2388. 84

Jin, S., Ji, C., Yan, C., and Xing, J. (2018). Tft-lcd mura defect detection using dct and the dual-γ piecewise exponential transform. *Precis. Eng.*, 54:371–378. 126

Jin, Y. W., Jia, S., Ashraf, A. B., and Hu, P. (2020). Integrative data augmentation with u-net segmentation masks improves detection of lymph node metastases in breast cancer patients. *Cancers*, 12(10):2934. 104

Joglekar-Javadekar, M., Van Laere, S., Bourne, M., Moalwi, M., Finetti, P., Vermeulen, P. B., Birnbaum, D., Dirix, L. Y., Ueno, N., Carter, M., et al. (2017). Characterization and targeting of platelet-derived growth factor receptor alpha (pdgfra) in inflammatory breast cancer (ibc). *Neoplasia*, 19(7):564–573. 43

Joshi, A., Mishra, G., and Sivaswamy, J. (2020). Explainable disease classification via weakly-supervised segmentation. In *Interpretable and Annotation-Efficient Learning for Medical Image Computing*, pages 54–62. Springer International Publishing. 105

Kalager, M., Adami, H.-O., Bretthauer, M., and Tamimi, R. M. (2012). Overdiagnosis of invasive breast cancer due to mammography screening: results from the norwegian screening program. *Annals of internal medicine*, 156(7):491–499. 46

Kang, S., Lee, J., Song, K., and Pahk, H. (2009). Automatic defect classification of tft-lcd panels using machine learning. In *2009 IEEE International Symposium on Industrial Electronics*, pages 2175–2177. IEEE. 59

Kaplun, D., Krasichkov, A., Chetyrbok, P., Oleinikov, N., Garg, A., and Pannu, H. S. (2021). Cancer cell profiling using image moments and neural networks with model agnostic explainability: A case study of breast cancer histopathological (BreakHis) database. *Mathematics*, 9(20):2616. 105

Kasraeian, S., Allison, D. C., Ahlmann, E. R., Fedenko, A. N., and Menendez, L. R. (2010). A comparison of fine-needle aspiration, core biopsy, and surgical biopsy in the diagnosis of extremity soft tissue masses. *Clinical Orthopaedics and Related Research®*, 468(11):2992–3002. 51

Khotanzad, A. and Hong, Y. (1990). Invariant image recognition by zernike moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5):489–497. 105

Kim, J.-H., Ahn, S., Jeon, J. W., and Byun, J.-E. (2001). A high-speed high-resolution vision system for the inspection of tft lcd. In *ISIE 2001. 2001 IEEE International Symposium on Industrial Electronics Proceedings (Cat. No. 01TH8570)*, volume 1, pages 101–105. IEEE. 58

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. 69

Komenaka, I. K., El-Tamer, M. B., Troxel, A., Hamele-Bena, D., Joseph, K.-A., Horowitz, E., Ditkoff, B.-A., and Schnabel, F. R. (2004). Pure mucinous carcinoma of the breast. *The American journal of surgery*, 187(4):528–532. 1, 33

Kossoff, M. B. (2000). Ultrasound of the breast. *World journal of surgery*, 24(2):143–157. xi, 9, 46, 47

Krithiga, R. and Geetha, P. (2020). Breast cancer detection, segmentation and classification on histopathology images analysis: A systematic review. *Archives of Computational Methods in Engineering*, 28(4):2607–2619. 104

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc. xi, xii, 12, 61, 62

Kuhl, C. K. (2019). Abbreviated magnetic resonance imaging (mri) for breast cancer screening: rationale, concept, and transfer to clinical practice. *Annu Rev Med*, 70(1):501–519. 2, 34

Kumar, A., Singh, S. K., Saxena, S., Lakshmanan, K., Sangaiah, A. K., Chauhan, H., Shrivastava, S., and Singh, R. K. (2020). Deep feature learning for histopathological image classification of canine mammary tumors and human breast cancer. *Information Sciences*, 508:405–421. 2, 34, 104

Le, X., Mei, J., Zhang, H., Zhou, B., and Xi, J. (2020). A learning-based approach for surface defect detection using small image datasets. *Neurocomputing*, 408:112–120. 126

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., and Jackel, L. (1989). Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2. 12, 61

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324. xi, xii, 12, 61, 62

Lee, K.-B., Ko, M.-S., Lee, J. J., Koo, T.-M., and Park, K.-h. (2004). Defect detection method for tft-lcd panel based on saliency map model. In *2004 IEEE Region 10 Conference TENCON 2004.*, pages 223–226. IEEE. 59

Lee, M., Jeon, J., and Lee, H. (2022). Explainable ai for domain experts: a post hoc analysis of deep learning for defect classification of tft–lcd panels. *J. Intell. Manuf.*, 33(6):1747–1759. 4, 35, 127

Lehman, C. D. and Schnall, M. D. (2005). Imaging in breast cancer: magnetic resonance imaging. *Breast Cancer Research*, 7(5):1–5. 49

Lei, H., Wang, B., Wu, H., and Wang, A. (2018a). Defect detection for polymeric polarizer based on faster r-cnn. *J. Inf. Hiding Multim. Signal Process.*, 9(6):1414–1420. 59

Lei, J., Gao, X., Feng, Z., Qiu, H., and Song, M. (2018b). Scale insensitive and focus driven mobile screen defect detection in industry. *Neurocomputing*, 294:72–81. 126

Lerousseau, M., Vakalopoulou, M., Classe, M., Adam, J., Battistella, E., Carré, A., Estienne, T., Henry, T., Deutsch, E., and Paragios, N. (2020). Weakly supervised multiple instance learning histopathological tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 470–479. Springer. 84

Li, C., Wang, X., Liu, W., and Latecki, L. J. (2018). DeepMitosis: Mitosis detection via deep detection, verification and segmentation networks. *Medical Image Analysis*, 45:121–133. 104

Li, C., Wang, X., Liu, W., Latecki, L. J., Wang, B., and Huang, J. (2019a). Weakly supervised mitosis detection in breast histopathology images using concentric loss. *Med. Image Anal.*, 53:165–178. 126

Li, J., Li, W., Sisk, A., Ye, H., Wallace, W. D., Speier, W., and Arnold, C. W. (2021a). A multi-resolution model for histopathology image classification and localization with multiple instance learning. *Computers in biology and medicine*, 131:104253. 84

Li, J. and Wang, H. (2022). Surface defect detection of vehicle light guide plates based on an improved retinanet. *Meas. Sci. Technol.*, 33(4):045401. 4, 35, 127

Li, J.-p., Zhang, X.-m., Zhang, Z., Zheng, L.-h., Jindal, S., and Liu, Y.-j. (2019b). Association of p53 expression with poor prognosis in patients with triple-negative breast invasive ductal carcinoma. *Medicine*, 98(18). 1, 33

Li, L., Wang, B., Verma, M., Nakashima, Y., Kawasaki, R., and Nagahara, H. (2021b). Scouter: Slot attention-based classifier for explainable image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1046–1055. 105, 110

Li, L., Zhou, T., Wang, W., Li, J., and Yang, Y. (2022). Deep hierarchical semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1246–1257. 16, 78

Li, Y. and Li, J. (2021). An end-to-end defect detection method for mobile phone light guide plate via multitask learning. *IEEE Trans. Instrum. Meas.*, 70:1–13. 4, 35, 127

Lian, J., Jia, W., Zareapoor, M., Zheng, Y., Luo, R., Jain, D. K., and Kumar, N. (2019). Deep-learning-based small surface defect detection via an exaggerated local variation-based generative adversarial network. *IEEE Transactions on Industrial Informatics*, 16(2):1343–1351. 16, 78

Liberman, L. and Menell, J. H. (2002). Breast imaging reporting and data system (bi-rads). *Radiologic Clinics*, 40(3):409–430. 52

Lilly, A. J., Johnson, M., Kuzmiak, C. M., Ollila, D. W., O'Connor, S. M., Hertel, J. D., and Calhoun, B. C. (2020). Mri-guided core needle biopsy of the breast: Radiology-pathology correlation and impact on clinical management. *Annals of diagnostic pathology*, 48:151563. 51

Lin, A., Chen, B., Xu, J., Zhang, Z., Lu, G., and Zhang, D. (2022). Ds-transunet: Dual swin transformer u-net for medical image segmentation. *IEEE Transactions on Instrumentation and Measurement*. 4, 16, 35, 67, 78

Liu, E., Chen, K., Xiang, Z., and Zhang, J. (2020a). Conductive particle detection via deep learning for acf bonding in tft-lcd manufacturing. *J. Intell. Manuf.*, 31(4):1037–1049. 4, 35, 127, 138

Liu, J., Wang, C., Su, H., Du, B., and Tao, D. (2019). Multistage gan for fabric defect detection. *IEEE Transactions on Image Processing*, 29:3388–3400. 16, 78

Liu, Y., Gao, Y., and Yin, W. (2020b). An improved analysis of stochastic gradient descent with momentum. *Advances in Neural Information Processing Systems*, 33:18261–18271. 69

Liu, Y.-H., Lin, S.-H., Hsueh, Y.-L., and Lee, M.-J. (2009). Automatic target defect identification for tft-lcd array process inspection using kernel fcm-based fuzzy svdd ensemble. *Expert Systems with Applications*, 36(2):1978–1998. 59

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440. xi, xii, 14, 16, 65, 71, 78, 134

Ma, N., Zhang, X., Zheng, H.-T., and Sun, J. (2018). Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131. 4, 16, 36, 78

Ma, Z. and Gong, J. (2019). An automatic detection method of mura defects for liquid crystal display. In *2019 Chinese Control Conference (CCC)*, pages 7722–7727. IEEE. 126

Mann, R. M., Kuhl, C. K., Kinkel, K., and Boetes, C. (2008). Breast mri: guidelines from the european society of breast imaging. *European radiology*, 18(7):1307–1318. 48

Marquez-Neila, P., Baumela, L., and Alvarez, L. (2013). A morphological approach to curvature-based evolution of curves and surfaces. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(1):2–17. 135

Masannat, Y. A., Bains, S. K., Pinder, S. E., and Purushotham, A. D. (2013). Challenges in the management of pleomorphic lobular carcinoma in situ of the breast. *The Breast*, 22(2):194–196. 1, 33

Mateo, A. M., Pezzi, T. A., Sundermeyer, M., Kelley, C. A., Klimberg, V. S., and Pezzi, C. M. (2017). Chemotherapy significantly improves survival for patients with t1c-t2n0m0 medullary breast cancer: 3739 cases from the national cancer data base. *Annals of surgical oncology*, 24(4):1050–1056. 1, 33

Meng, Z., Zhao, Z., Li, B., Su, F., and Guo, L. (2021). A cervical histopathology dataset for computer aided diagnosis of precancerous lesions. *IEEE Transactions on Medical Imaging*, 40(6):1531–1541. 2, 34

Moon, D. H., Maddahi, J., Silverman, D. H., Glaspy, J. A., Phelps, M. E., and Hoh, C. K. (1998). Accuracy of whole-body fluorine-18-fdg pet for the detection of recurrent or metastatic breast carcinoma. *Journal of Nuclear Medicine*, 39(3):431–435. 50

Mukherjee, A., Chaudhuri, S., Dutta, P. K., Sen, S., and Patra, A. (2006). An object-based coding scheme for frontal surface of defective fluted ingot. *ISA Trans.*, 45(1):1–8. 126

Nakashima, K. (1994). Hybrid inspection system for lcd color filter panels. In *Conference Proceedings. 10th Anniversary. IMTC/94. Advanced Technologies in I & M. 1994 IEEE Instrumentation and Measurement Technolgy Conference (Cat. No. 94CH3424-9)*, pages 689–692. IEEE. 58

Ng, H.-F. (2006). Automatic thresholding for defect detection. *Pattern Recognit. Lett.*, 27(14):1644–1649. 126, 135

Nguyen, D.-K., Ju, J., Booij, O., Oswald, M. R., and Snoek, C. G. (2022). Boxer: Box-attention for 2d and 3d transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4773–4782. 16, 78

Noh, C.-H., Lee, S.-L., Kim, D.-H., and Chung, C.-W. (2009). Effective defect classification for flat display panel film images. In *Proceedings of the 2009 International Conference on Hybrid Information Technology*, pages 264–267. 59

O'Flynn, E., Wilson, A., and Michell, M. (2010). Image-guided breast biopsy: state-of-the-art. *Clinical radiology*, 65(4):259–270. 52

Onega, T., Goldman, L. E., Walker, R. L., Miglioretti, D. L., Buist, D. S., Taplin, S., Geller, B. M., Hill, D. A., and Smith-Bindman, R. (2016). Facility mammography volume in relation to breast cancer screening outcomes. *Journal of medical screening*, 23(1):31–37. 2, 34

O'Shea, R. J., Horst, C., Manickavasagar, T., Hughes, D., Cusack, J., Tsoka, S., Cook, G., and Goh, V. (2022). Weakly supervised unet: an image classifier which learns to explain itself. *bioRxiv*. 105

Ouyang, W., Zeng, X., Wang, X., Qiu, S., Luo, P., Tian, Y., Li, H., Yang, S., Wang, Z., Li, H., et al. (2016). Deepid-net: Object detection with deformable part based convolutional neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(7):1320–1334. 126

Ozmen, N., Dapp, R., Zapf, M., Gemmeke, H., Ruiter, N. V., and van Dongen, K. W. (2015). Comparing different ultrasound imaging methods for breast cancer detection. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, 62(4):637–646. 2, 33, 47

PARK, H.-M., LIM, H.-S., KI, S.-Y., Lee, H.-j., LEE, J.-S., and PARK, M.-H. (2021). Invasive ductal carcinoma originating from a borderline phyllodes tumor in a young female: A case report. *Journal of the Korean Radiological Society*, pages 971–976. 1, 33

Patil, A., Tamboli, D., Meena, S., Anand, D., and Sethi, A. (2019). Breast cancer histopathology image classification and localization using multiple instance learning. In *2019 IEEE International WIE conference on electrical and computer engineering (WIECON-ECE)*, pages 1–4. IEEE. xii, 73, 74, 89, 90, 105, 113, 114

Pelosi, E., Messa, C., Sironi, S., Picchio, M., Landoni, C., Bettinardi, V., Gianolli, L., Del Maschio, A., Gilardi, M. C., and Fazio, F. (2004). Value of integrated pet/ct for lesion localisation in cancer patients: a comparative study. *European Journal of Nuclear Medicine and Molecular Imaging*, 31(7):932–939. 50

Petsiuk, V., Jain, R., Manjunatha, V., Morariu, V. I., Mehra, A., Ordonez, V., and Saenko, K. (2021). Black-box explanation of object detectors via saliency maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11443–11452. 105, 110

Pohle, R. and Toennies, K. D. (2001). Segmentation of medical images using adaptive region growing. In *Medical Imaging 2001: Image Processing*, volume 4322, pages 1337–1346. SPIE. 135

Qi, Q., Li, Y., Wang, J., Zheng, H., Huang, Y., Ding, X., and Rohde, G. K. (2019). Label-efficient breast cancer histopathological image classification. *IEEE Journal of Biomedical and Health Informatics*, 23(5):2108–2116. 89, 90, 118

Qiu, Y., Yan, S., Gundreddy, R. R., Wang, Y., Cheng, S., Liu, H., and Zheng, B. (2017). A new approach to develop computer-aided diagnosis scheme of breast mass classification using deep learning technology. *Journal of X-ray Science and Technology*, 25(5):751–763. 2, 34

Quellec, G., Cazuguel, G., Cochener, B., and Lamard, M. (2017). Multiple-instance learning for medical image and video analysis. *IEEE reviews in biomedical engineering*, 10:213–234. 83

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR. 104

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (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. 16, 78

Redmon, J. and Farhadi, A. (2017). Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271. 16, 78

Redmon, J. and Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*. 16, 78

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28. 16, 78

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?". In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 105

Rodriguez-Sampaio, M., Rincón, M., Valladares-Rodriguez, S., and Bachiller-Mayoral, M. (2022). Explainable artificial intelligence to detect breast cancer: A qualitative case-based visual interpretability approach. In *Artificial Intelligence in Neuroscience: Affective Analysis and Health Applications*, pages 557–566. Springer International Publishing. 105

Roganovic, D., Djilas, D., Vujnovic, S., Pavic, D., and Stojanov, D. (2015). Breast mri, digital mammography and breast tomosynthesis: comparison of three methods for early detection of breast cancer. *Bosnian journal of basic medical sciences*, 15(4):64. 2, 34

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer. xii, 4, 16, 35, 66, 78, 98, 99

Rony, J., Belharbi, S., Dolz, J., Ayed, I. B., McCaffrey, L., and Granger, E. (2019). Deep weakly-supervised learning methods for classification and localization in histology images: a survey. *arXiv preprint arXiv:1909.03354*. 9, 52, 84

Ruan, L., Gao, B., Wu, S., and Woo, W. L. (2020). Deftectnet: Joint loss structured deep adversarial network for thermography defect detecting system. *Neurocomputing*, 417:441–457. 127

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*. 69

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533–536. 61

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252. 71

Sakorafas, G., Blanchard, K., Sarr, M., and Farley, D. (2001). Paget's disease of the breast. *Cancer treatment reviews*, 27(1):9–18. 1, 33

Samala, R. K., Chan, H.-P., Hadjiiski, L., Helvie, M. A., Richter, C. D., and Cha, K. H. (2018). Breast cancer diagnosis in digital breast tomosynthesis: effects of training sample size on multi-stage transfer learning using deep neural nets. *IEEE transactions on medical imaging*, 38(3):686–696. 88

Scheidhauer, K., Scharl, A., Pietrzyk, U., Wagner, R., Göhring, U.-J., Schomäcker, K., and Schicha, H. (1996). Qualitative [18f] fdg positron emission tomography in primary breast cancer: clinical relevance and practicability. *European journal of nuclear medicine*, 23(6):618–623. 50

Schirris, Y., Gavves, E., Nederlof, I., Horlings, H. M., and Teuwen, J. (2022). Deepsmile: Contrastive self-supervised pre-training benefits msi and hrd classification directly from h&e whole-slide images in colorectal and breast cancer. *Medical Image Analysis*, 79:102464. 2, 34, 104

Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., and Rueckert, D. (2019). Attention gated networks: Learning to leverage salient regions in medical images. *Med. Image Anal.*, 53:197–207. xii, 4, 16, 21, 26, 35, 67, 78, 80, 98, 99, 128, 132, 138

Schmitz, R., Madesta, F., Nielsen, M., Krause, J., Steurer, S., Werner, R., and Rösch, T. (2021). Multi-scale fully convolutional neural networks for histopathology image segmentation: From nuclear aberrations to the global tissue architecture. *Medical Image Analysis*, 70:101996. 104

Schnall, M. D. (2000). Breast imaging technology application of magnetic resonance imaging to early detection of breast cancer. *Breast Cancer Research*, 3(1):1–5. 49

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017a). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. xii, 71, 72, 75, 105, 113, 114

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017b). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626. 126

Shaham, T. R., Dekel, T., and Michaeli, T. (2019). Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4570–4580. 104

Shahid, H., Wiedenhoefer, J. F., Carol Dornbluth MD, D., Pamela Otto, M., Kist, K. A., et al. (2016). An overview of breast mri. *Applied Radiology*, 45(10):7. xi, 48, 49

Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al. (2021). Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34:2136–2147. 84

Sharma, Y., Shrivastava, A., Ehsan, L., Moskaluk, C. A., Syed, S., and Brown, D. (2021). Cluster-to-conquer: A framework for end-to-end multi-instance learning for whole slide image classification. In *Medical Imaging with Deep Learning*, pages 682–698. PMLR. 84

Shen, Y., Wu, N., Phang, J., Park, J., Liu, K., Tyagi, S., Heacock, L., Kim, S. G., Moy, L., Cho, K., and Geras, K. J. (2021). An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *Medical Image Analysis*, 68:101908. 105

Silverstein, M. J., Lewinsky, B. S., Waisman, J. R., Gierson, E. D., Colburn, W. J., Senofsky, G. M., and Gamagami, P. (1994). Infiltrating lobular carcinoma. is it different from infiltrating duct carcinoma? *Cancer*, 73(6):1673–1677. 41

Simonyan, K. and Zisserman, A. (2014a). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. 2, 4, 12, 16, 18, 34, 36, 62, 78, 79, 86, 126, 130

Simonyan, K. and Zisserman, A. (2014b). Very deep convolutional networks for large-scale image recognition. 24, 73, 107, 118

Singh, A., Sengupta, S., and Lakshminarayanan, V. (2020). Explainable deep learning models in medical image analysis. *Journal of Imaging*, 6(6):52. 105

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. 105

Song, G., Song, K., and Yan, Y. (2020). Edrnet: Encoder–decoder residual network for salient object detection of strip steel surface defects. *IEEE Trans. Instrum. Meas.*, 69(12):9709–9719. 4, 16, 35, 36, 78, 127, 138

Song, Y., Chang, H., Huang, H., and Cai, W. (2017). Supervised intra-embedding of fisher vectors for histopathology image classification. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*, pages 99–106. Springer International Publishing. 2, 34, 104, 118

Soo, M. S., Shelby, R. A., and Johnson, K. S. (2019). Optimizing the patient experience during breast biopsy. *Journal of Breast Imaging*, 1(2):131–138. 2, 34

Spanhol, F. A., Oliveira, L. S., Petitjean, C., and Heutte, L. (2016a). Breast cancer histopathological image classification using convolutional neural networks. In *2016 international joint conference on neural networks (IJCNN)*, pages 2560–2567. IEEE. 2, 34, 89, 90, 104, 118

Spanhol, F. A., Oliveira, L. S., Petitjean, C., and Heutte, L. (2016b). A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63(7):1455–1462. 10, 52, 112

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. 75, 105

Sree, S. V., Ng, E. Y.-K., Acharya, R. U., and Faust, O. (2011). Breast imaging: a survey. *World journal of clinical oncology*, 2(4):171. 46

Stingl, J., Raouf, A., Eirew, P., and Eaves, C. J. (2006). Deciphering the mammary epithelial cell hierarchy. *Cell cycle*, 5(14):1519–1522. 7, 41

Stingl, J., Raouf, A., Emerman, J. T., and Eaves, C. J. (2005). Epithelial progenitors in the normal human mammary gland. *Journal of mammary gland biology and neoplasia*, 10(1):49–59. 7, 41

Strayer, D. S. (2015). *Rubin's pathology: clinicopathologic foundations of medicine*. Jefferson Faculty Books. 104

Su, X., He, Z., and Ma, P. (2008). An automatic detection algorithm for tft-lcd micro display defects. *Journal of Harbin Institute of Technology*, 40(11):1756–1760. 58

Suarez, M., Perez-Castejon, M., Jimenez, A., Domper, M., et al. (2002). Early diagnosis of recurrent breast cancer with edg-pet in patients with progressive elevation of serum tumor markers. *The Quarterly Journal of Nuclear Medicine and Molecular Imaging*, 46(2):113. 50

Sudharshan, P., Petitjean, C., Spanhol, F., Oliveira, L. E., Heutte, L., and Honeine, P. (2019). Multiple instance learning for histopathological breast cancer image classification. *Expert Systems with Applications*, 117:103–111. 2, 34, 84, 89, 90, 104, 118

Sun, Y., Huang, X., Wang, Y., Zhou, H., and Zhang, Q. (2021). Magnification-independent histopathological image classification with similarity-based multi-scale embeddings. *arXiv preprint arXiv:2107.01063*. 2, 34, 89, 90, 104, 118

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR. 105

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249. 1, 33, 104

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*. 4, 13, 16, 36, 63, 78

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015a). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9. xi, xii, 4, 13, 16, 36, 63, 78

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015b). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 73, 109

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826. 4, 13, 16, 36, 63, 78

Tan, M. and Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR. 105

Tan, P. H., Tse, G., and Bay, B. H. (2008). Mucinous breast lesions: diagnostic challenges. *Journal of clinical pathology*, 61(1):11–19. 42

Tanis, P. J., Nieweg, O. E., Olmos, R. A. V., and Kroon, B. B. (2001). Anatomy and physiology of lymphatic drainage of the breast from the perspective of sentinel node biopsy. *Journal of the American College of Surgeons*, 192(3):399–409. 7, 41

Teed, Z. and Deng, J. (2020). RAFT: Recurrent all-pairs field transforms for optical flow. In *Computer Vision – ECCV 2020*, pages 402–419. Springer International Publishing. 104

Tey, W. K., Kuang, Y. C., Ooi, M. P.-L., and Khoo, J. J. (2018). Automated quantification of renal interstitial fibrosis for computer-aided diagnosis: A comprehensive tissue structure segmentation method. *Computer Methods and Programs in Biomedicine*, 155:109–120. 104

Thomas, E., Pawan, S., Kumar, S., Horo, A., Niyas, S., Vinayagamani, S., Kesavadas, C., and Rajan, J. (2020). Multi-res-attention unet: a cnn model for the segmentation of focal cortical dysplasia lesions from magnetic resonance images. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1724–1734. 4, 16, 35, 67, 78

Tian, C., Fei, L., Zheng, W., Xu, Y., Zuo, W., and Lin, C.-W. (2020). Deep learning on image denoising: An overview. *Neural Networks*, 131:251–275. 2, 34

Tong, L., Wong, W. K., and Kwong, C. K. (2016). Differential evolution-based optimal gabor filter model for fabric inspection. *Neurocomputing*, 173:1386–1401. 3, 35, 126

Tsai, D.-M., Chiang, I.-Y., and Tsai, Y.-H. (2011). A shift-tolerant dissimilarity measure for surface defect detection. *IEEE Trans. Ind. Informat.*, 8(1):128–137. 3, 35, 126

Tsai, D.-M., Chuang, S.-T., and Tseng, Y.-H. (2007). One-dimensional-based automatic defect inspection of multiple patterned tft-lcd panels using fourier image reconstruction. *Int. J. Prod. Res.*, 45(6):1297–1321. 3, 35, 58, 126

Tsai, D.-M. and Hung, C.-Y. (2005). Automatic defect inspection of patterned thin film transistor-liquid crystal display (tft-lcd) panels using one-dimensional fourier reconstruction and wavelet decomposition. *Int. J. Prod. Res.*, 43(21):4589–4607. 3, 35, 58, 126

Tsai, D.-M. and Lai, S.-C. (2008). Defect detection in periodically patterned surfaces using independent component analysis. *Pattern Recognition*, 41(9):2812–2832. 58

Turner, N. C. and Reis-Filho, J. S. (2006). Basal-like breast cancer and the brca1 phenotype. *Oncogene*, 25(43):5846–5853. 44

Van der Putte, S., Toonstra, J., and Hennipman, A. (1995). Mammary paget's disease confined to the areola and associated with multifocal toker cell hyperplasia. *The American journal of dermatopathology*, 17(5):487–493. 43

van Rijthoven, M., Balkenhol, M., Siliņa, K., van der Laak, J., and Ciompi, F. (2021). HookNet: Multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images. *Medical Image Analysis*, 68:101890. 104

Van Schoor, G., Moss, S., Otten, J., Donders, R., Paap, E., Den Heeten, G., Holland, R., Broeders, M., and Verbeek, A. (2011). Increasingly strong reduction in breast cancer mortality due to screening. *British journal of cancer*, 104(6):910–914. 46

Vasilic, S. and Hocenski, Z. (2006). The edge detecting methods in ceramic tiles defects detection. In *2006 IEEE International Symposium on Industrial Electronics*, volume 1, pages 469–472. 126, 135

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30. 31, 153

Vercher-Conejero, J. L., Pelegrí-Martinez, L., Lopez-Aznar, D., and Cózar-Santiago, M. D. P. (2015). Positron emission tomography in breast cancer. *Diagnostics*, 5(1):61–83. 2, 34

Voulodimos, A., Doulamis, N., Doulamis, A., and Protopapadakis, E. (2018). Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018. 2, 34

Vu, T., Lai, P., Raich, R., Pham, A., Fern, X. Z., and Rao, U. A. (2020). A novel attribute-based symmetric multiple instance learning for histopathological image analysis. *IEEE transactions on medical imaging*, 39(10):3125–3136. 84

Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M. (2022). Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*. 16, 78

Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., and Hu, X. (2020). Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 105, 110, 113, 114

Wang, L. (2017). Early diagnosis of breast cancer. *Sensors*, 17(7):1572. 47

Wang, P., Li, Z., and Pei, Y. (2018). In situ high temperature microwave microscope for nondestructive detection of surface and sub-surface defects. *Opt. Express*, 26(8):9595–9606. 3, 35, 126

Weedon-Fekjær, H., Li, X., and Lee, S. (2021). Estimating the natural progression of non-invasive ductal carcinoma in situ breast cancer lesions using screening data. *Journal of Medical Screening*, 28(3):302–310. 1, 33

Weigelt, B., Horlings, H., Kreike, B., Hayes, M., Hauptmann, M., Wessels, L., De Jong, D., Van de Vijver, M., Veer, L. V., and Peterse, J. (2008). Refinement of breast cancer classification by molecular characterization of histological special types. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, 216(2):141–150. 42

West, A.-K. V., Wullkopf, L., Christensen, A., Leijnse, N., Tarp, J. M., Mathiesen, J., Erler, J. T., and Oddershede, L. B. (2017). Division induced dynamics in non-invasive and invasive breast cancer. *Biophysical Journal*, 112(3):123a. 1, 33

Wikimedia, C. (2019). File:mammo breast cancer.jpg — wikimedia commons, the free media repository. [Online; accessed 9-January-2023]. xi, 45

Wikimedia, C. (2020). File:histopathology of invasive lobular carcinoma, next to lobular carcinoma in situ, annotated.jpg — wikimedia commons, the free media repository. [Online; accessed 9-January-2023]. xii, 52

Wikimedia, C. (2021). File:blausen 0628 mammogram.png — wikimedia commons, the free media repository. [Online; accessed 9-January-2023]. xi, 45

Wilson, A. C., Roelofs, R., Stern, M., Srebro, N., and Recht, B. (2017). The marginal value of adaptive gradient methods in machine learning. *Advances in neural information processing systems*, 30. 69

Xu, B., Liu, J., Hou, X., Liu, B., Garibaldi, J., Ellis, I. O., Green, A., Shen, L., and Qiu, G. (2019a). Attention by selection: A deep selective attention approach to breast cancer classification. *IEEE transactions on medical imaging*, 39(6):1930–1941. 2, 16, 34, 78

Xu, G., Song, Z., Sun, Z., Ku, C., Yang, Z., Liu, C., Wang, S., Ma, J., and Xu, W. (2019b). Camel: A weakly supervised learning framework for histopathology image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10682–10691. 84

Yanase, J. and Triantaphyllou, E. (2019). A systematic survey of computer-aided diagnosis in medicine: Past and present developments. *Expert Systems with Applications*, 138:112821. 104

Yao, J. and Li, J. (2022). Ayolov3-tiny: An improved convolutional neural network architecture for real-time defect detection of pad light guide plates. *Comput. Ind.*, 136:103588. 4, 35, 127

Yedjou, C. G., Tchounwou, S. S., Grigsby, J., Johnson, K., and Tchounwou, P. B. (2022). Improving invasive breast cancer care using machine learning technology. *Journal ISSN*, 2766:2276. 1, 33

Yeh, E. D., Jacene, H. A., Bellon, J. R., Nakhlis, F., Birdwell, R. L., Georgian-Smith, D., Giess, C. S., Hirshfield-Bartek, J., Overmoyer, B., and Van den Abbeele, A. D. (2013). What radiologists need to know about diagnosis and treatment of inflammatory breast cancer: a multidisciplinary approach. *Radiographics*, 33(7):2003–2017. 43

Yuan, X., Wu, L., and Peng, Q. (2015). An improved otsu method using the weighted object variance for defect detection. *Appl. Surface Sci.*, 349:472–484. 3, 35, 126

Zangheri, B., Messa, C., Picchio, M., Gianolli, L., Landoni, C., and Fazio, F. (2004). Pet/ct and breast cancer. *European journal of nuclear medicine and molecular imaging*, 31(1):S135–S142. xii, 50

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*. 69

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014*, pages 818–833. Springer International Publishing. 105

Zhang, T., Lin, G., Liu, W., Cai, J., and Kot, A. (2020). Splitting vs. merging: Mining object regions with discrepancy and intersection loss for weakly supervised semantic segmentation. In *Computer Vision – ECCV 2020*, pages 663–679. Springer International Publishing. 104

Zhang, T., Lu, R., and Zhang, S. (2016). Surface defect inspection of tft-lcd panels based on 2d dft. *Opto-Electronic Engineering*, 43(3):7–15. 59

Zhang, X., Wei, Y., Feng, J., Yang, Y., and Huang, T. S. (2018a). Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1325–1334. xii, 2, 34, 73, 74, 105, 113, 114

Zhang, X., Zhou, X., Lin, M., and Sun, J. (2018b). Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856. 4, 36, 104

Zhang, Y. and Kleer, C. G. (2016). Phyllodes tumor of the breast: histopathologic features, differential diagnosis, and molecular/genetic updates. *Archives of pathology & laboratory medicine*, 140(7):665–671. 43

Zhang, Y. and Zhang, J. (2005). A fuzzy neural network approach for quantitative evaluation of mura in tft-lcd. In *2005 International Conference on Neural Networks and Brain*, volume 1, pages 424–427. IEEE. 59

Zhao, Y., Yang, F., Fang, Y., Liu, H., Zhou, N., Zhang, J., Sun, J., Yang, S., Menze, B., Fan, X., et al. (2020). Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4837–4846. 84

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P. H., et al. (2021). Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890. 2, 34

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. xii, 71, 105

Zhou, T., Wang, W., Konukoglu, E., and Van Gool, L. (2022). Rethinking semantic segmentation: A prototype view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2582–2593. 16, 78

Zhou, X., Zheng, Z., Li, Y., Zhao, W., Lin, Y., Zhang, J., and Sun, Q. (2021). The clinical features and prognosis of patients with mucinous breast carcinoma compared with those with infiltrating ductal carcinoma: a population-based study. *BMC cancer*, 21(1):1–9. 1, 33

Zhou, Z.-H. (2004). Multi-instance learning: A survey. *Department of Computer Science & Technology, Nanjing University, Tech. Rep*, 1. 83

Zhu, C., Song, F., Wang, Y., Dong, H., Guo, Y., and Liu, J. (2019). Breast cancer histopathology image classification through assembling multiple compact CNNs. *BMC Medical Informatics and Decision Making*, 19(1). 118

Zou, F., Shen, L., Jie, Z., Zhang, W., and Liu, W. (2019). A sufficient condition for convergences of adam and rmsprop. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11127–11135. 69

Zou, Q., Zhang, Z., Li, Q., Qi, X., Wang, Q., and Wang, S. (2018). Deepcrack: Learning hierarchical convolutional features for crack detection. *IEEE Trans. Image Process.*, 28(3):1498–1512. 4, 16, 35, 36, 78, 127, 138

Zunair, H. and Hamza, A. B. (2021). Sharp u-net: depthwise convolutional network for biomedical image segmentation. *Computers in Biology and Medicine*, 136:104699. 4, 16, 35, 67, 78

Zuo, Q., Chen, S., and Wang, Z. (2021). R2au-net: attention recurrent residual convolutional neural network for multimodal medical image segmentation. *Security and Communication Networks*, 2021. 67, 98, 99

FOLIO ADMINISTRATIF

THESE DE L'INSA LYON, MEMBRE DE L'UNIVERSITE DE LYON

NOM : HE
(avec précision du nom de jeune fille, le cas échéant)

DATE de SOUTENANCE : 29/03/2023

Prénoms : Feng

TITRE : Configurable Convolutional Neural Networks: Applications to Breast Cancer Explainable Classification and Display Panel Defect Detection

NATURE : Doctorat

Numéro d'ordre : 2023ISAL0022

Ecole doctorale : Electronique, Électrotechnique, Automatique (EEA) – ED160

Spécialité : Traitement du Signal et de l'Image

RESUME :
Les méthodes actuelles d'apprentissage profond, telles que les réseaux de neurones convolutifs (CNNs), sont souvent dédiées à une tâche et à un objet spécifiques ; leur architecture de réseau est généralement fixe, ce qui limite leur généralisabilité et les empêche d'aborder de multiples scénarios avec des objectifs différents. Pour réaliser à la fois la classification explicable du cancer du sein et la détection en ligne des défauts des panneaux d'affichage, nous proposons un réseau de neurones convolutif configurable (ConfigNet) capable d'être transformé en différentes configurations selon les tâches et les objets en question. Le ConfigNet présente deux configurations fonctionnelles principales. La première est composée d'un module d'extraction de caractéristiques (FEM), d'un générateur de cartes de décision (DMG) et d'un classificateur ; elle est consacrée à la classification explicative d'images, pour laquelle nous proposons deux structures DMG et un classificateur de mise en commun des moyennes pondérées (WAP) pour les images histopathologiques du cancer du sein. La seconde est une configuration codeur-décodeur consacrée à la segmentation et à la localisation d'objets. Dans cette deuxième configuration, nous proposons un décodeur favorisant l'efficacité et un module de fusion de caractéristiques par éléments (EFFM) guidant la connexion par saut entre l'encodeur et le décodeur pour la détection en ligne des défauts des panneaux d'affichage. En outre, nous développons un module de fusion de caractéristiques guidé par l'attention spatiale et l'attention de canal (SCAFFM) et un décodeur avec une structure de goulot d'étranglement pour la segmentation des tumeurs du sein. Le FEM ou le codeur dans ces deux configurations est construit par apprentissage par transfert à partir de CNNs existants ayant des couches convolutionnelles profondes.

MOTS-CLÉS : Apprentissage profond, Réseau neuronal convolutif configurable, Apprentissage faiblement supervisé, Classification explicable, XAI, Segmentation, Cancer du sein, Images histopathologiques, Détection des défauts, Panneau d'affichage

Laboratoire (s) de recherche : CREATIS

Directeur de thèse: ZHU Yuemin

Président de jury :

Composition du jury : VINCENT Nicole, YANG Jie, DUPONT Florent, LIU Wanyu, LIU Zhengjun, ZHU Yuemin