Real-Time Gastric Intestinal Metaplasia Semantic Segmentation with Multiple Abnormalities Using Deep Learning Approach

Mr. Passin Pornvoraphat

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering in Computer Engineering
Department of Computer Engineering
FACULTY OF ENGINEERING
Chulalongkorn University
Academic Year 2022

Copyright of Chulalongkorn University

การแบ่งส่วนรูปภาพชิ้นเนื้อชนิดเซลล์แบ่งตัวแบบผิดปกติในกระเพาะอาหารแบบทันทีพร้อมด้วยความผิดปกติที่หลากหลายโดยใช้กระบวนการการเรียนรู้เชิงลึก

นายพสิณ พรวรภัตซ์

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2565

ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

Thesis Title	Real-Time Gastric Intestinal Metaplasia Semantic Segmentation with Multiple Abnormalities Using Deep Learning Approach
By	Mr. Passin Pornvoraphat
Field of Study	Computer Engineering
Thesis Advisor	Associate Professor PEERAPON VATEEKUL, Ph.D.
Thesis Co Advisor	Kasenee Tiankanon, M.D.

Accepted by the FACULTY OF ENGINEERING, Chulalongkorn University in Partial Fulfillment of the Requirement for the Master of Engineering

.... Dean of the FACULTY OF ENGINEERING
(Professor SUPOT TEACHAVORASINSKUN, D.Eng.)

THESIS COMMITTEE

.... Chairman
(Professor BOONSERM KIJSIRIKUL, Ph.D.)

.... Thesis Advisor
(Associate Professor PEERAPON VATEEKUL, Ph.D.)

.... Thesis Co-Advisor
(Kasenee Tiankanon, M.D.)

.... Examiner
(PUNNARAI SIRICHAROEN, Ph.D.)

.... External Examiner
(Thanapat Kangkachit, Ph.D.)

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

พสิณ พรววรภัตช์ : การแบ่งส่วนรูปภาพชิ้นเนื้อชนิดเซลล์แบ่งตัวแบบผิดปกติในกระเพาะอาหารแบบทันทีพร้อมด้วยความผิดปกติที่หลากหลายโดยใช้กระบวนการการเรียนรู้เชิงลึก. (Real-Time Gastric Intestinal Metaplasia Semantic Segmentation with Multiple Abnormalities Using Deep Learning Approach) อ.ที่ปรึกษาหลัก : รศ. ดร.พีรพล เวทีกูล, อ.ที่ปรึกษาร่วม : พญ.เกศินี เธียรกานนท์

วิทยานิพนธ์นี้นำเสนอการแบ่งส่วนรูปภาพชิ้นเนื้อชนิดเซลล์แบ่งตัวแบบผิดปกติในกระเพาะอาหารแบบทันที (GIM) เมื่อเร็วๆ นี้ได้มีงานวิจัยที่การดำเนินการแบ่งส่วน GIM บนภาพส่องกล้องออกจากกระเพาะอาหารปกติ อย่างไรก็ตามการตรวจจับความผิดปกติในกระเพาะอาหารแบบทันทีนั้นยังมีข้อจำกัดอยู่และยังมีความท้าทายบางอย่างที่ยังไม่ได้รับการเติมเต็มเช่น ความหลายสีของโหมดสี (การส่องกล้องด้วยแสงสีขาวและการส่องกล้องด้วยแสงพิเศษ narrow-band imaging) รวมไปถึงความผิดปกติอื่นๆ (แผลสึกกร่อนและแผลพุพอง) และสุดท้ายความคลาดเคลื่อนเนื่องจากผลเฉลยที่ไม่แน่นอน ในที่นี้แบบจำลองที่เราใช้ BiSeNet สามารถแก้ปัญหาต่างๆ เกี่ยวกับ GIM ได้ ผลจากการใช้ auxiliary head และ additional loss ได้รับการพิสูจน์ว่าสามารถเพิ่มประสิทธิภาพในการตรวจจับ GIM อีกทั้งยังรองรับโทนสีที่หลากหลาย นอกจากนี้ได้มีการใช้เทคนิค "location-wise negative sampling" "jigsaw augmentation" และ "label smoothing" ในการเพิ่มประสิทธิภาพของ AI ปิดท้ายด้วยการใช้ threshold ที่แยกกันระหว่างภาพ NBI และ WLE เพื่อเพิ่มประสิทธิภาพของโมเดลขึ้นไปอีก งานวิจัยนี้ได้ดำเนินการ ณ ศูนย์ส่องกล้องโรงพยาบาลจุฬาลงกรณ์ โดยมีข้อมูล GIM จากการส่องกล้องที่ได้รับการพิสูจน์แล้ว 940 ภาพ และรูปภาพที่ไม่ใช่ GIM 1,239 ภาพ จากการไปทดสอบความสามารถของ AI พบว่า AI สามารถทำความเร็วถึง 173 เฟรมต่อวินาที ในแง่ศักยภาพ โมเดลเราสามารถเอาชนะ โมเดลมาตรฐานทั้งหมด และสามารถทำคะแนนประสิทธิภาพได้ดังต่อไปนี้ "F1-score" "sensitivity" "specificity" "positive predictive" "negative predictive" "accuracy" และ "IoU" ได้คะแนน 91%, 91%, 96%, 94% และ 55% ตามลำดับ สุดท้ายนี้วิธีการต่างๆ ที่ได้นำเสนอได้รับการทดสอบบนโมเดลมาตรฐานอื่นๆ เช่น STDC2-Seg50 และ BlazeNeo โดยได้คะแนน F1-score และ IoU อยู่ที่ 93% และ 56% สำหรับ STDC2-Seg50 และ 93% และ 56% สำหรับ BlazeNeo

สาขาวิชา	วิศวกรรมคอมพิวเตอร์	ลายมือชื่อนิสิต ....
ปีการศึกษา	2565	ลายมือชื่อ อ.ที่ปรึกษาหลัก ....
		ลายมือชื่อ อ.ที่ปรึกษาร่วม ....

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

6470420021 : MAJOR COMPUTER ENGINEERING

KEYWORD: Deep learning for vision, Gastric intestinal metaplasia, Real-time semantic segmentation, Upper gastrointestinal endoscopy

Passin Pornvoraphat : Real-Time Gastric Intestinal Metaplasia Semantic Segmentation with Multiple Abnormalities Using Deep Learning Approach. Advisor: Assoc. Prof. PEERAPON VATEEKUL, Ph.D. Co-advisor: Kasenee Tiankanon, M.D.

This thesis declares the segmentation of gastric intestinal metaplasia (GIM) in real-time. Recently, GIM segmentation of endoscopic images has been conducted to distinguish GIM from a healthy stomach. However, achieving real-time detection is difficult. Challenging conditions include multiple color modes (white light endoscopy and narrow-band imaging), other abnormal lesions (erosion and ulcer), noisy labels, etc. Herein, our model is based on BiSeNet and can overcome the many issues regarding GIM. Applying auxiliary head and loss boosts the performance on multiple color modes. In addition, pre-processing techniques, including location-wise negative sampling, jigsaw augmentation, and label smoothing, are utilized to improve detection performance. Finally, the decision threshold can be independently altered for each color mode. Work undertaken at King Chulalongkorn Memorial Hospital examined 940 histologically proven GIM images and 1239 non-GIM images, obtained over 173 FPS. In terms of accuracy, our model outperforms all baselines. Our results reveal F1-score, sensitivity, specificity, accuracy, and mean intersection over union (IoU), achieving 91%, 91%, 96%, 94%, and 55%, respectively. In addition, the effectiveness of the proposed methods was validated on baseline models, achieving F1-score and IoU values of 93% and 56% for STDC2-Seg50 and 93% and 56% for BlazeNeo.

Field of Study:	Computer Engineering	Student's Signature ....
Academic Year:	2022	Advisor's Signature ....
		Co-advisor's Signature ....

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ACKNOWLEDGEMENTS

I would like to express my gratitude to the great supervisor, Assoc. Prof. Dr. Peerapon Vateekul, for his dedication and support throughout the research. His guidance and feedbacks were invaluable in aiding me to accomplish any challenges and obstacles that came up along the way. Moreover, my thankfulness extends to the National Research Council of Thailand, Chulalongkorn University, and the Center of Excellence in Gastrointestinal Oncology, Chulalongkorn University, for their grant and support of my research. And I also appreciate the proficient endoscopists, namely Dr. Kasenee Tiankanon, Dr. Rapat Pittayanon, and Dr. Rungsun Rerknimitr, for their contribution to providing data, location usage, and valuable resources. Their guidance and profound comments about upper endoscopy were worthwhile and enabled us to conduct excellent research, especially with Dr. Kasenee Tiankanon. Her support helped our breakthrough study go smoothly and allowed us to publish in Elsevier. Mr. Phanukorn Sunthornwetchapong also deserves my gratitude for his contribution to deploying our AI model and carrying on with my project at the Center of Excellence. Again, my thank is given to Mr.Pasit Jakkrawankul for his continual and constructive comments on many parts of our proposed method. Last but not least, my family and my girlfriend are also the curial part of my life during my studies via financial and mental support. Remember my friends, Mr. Passakron Phuangthongkham and Mr. Mam Sothornin, etc., for being good friends throughout all my academic years. Finally, this thesis was possible with the facility of the Data Mind Laboratory, especially the high-end GPU: RTX 2080Ti.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Passin Pornvoraphat

TABLE OF CONTENTS

Page

.... iii

ABSTRACT (THAI) iii

.... iv

ABSTRACT (ENGLISH) iv

ACKNOWLEDGEMENTS v

TABLE OF CONTENTS vi

LIST OF TABLES x

LIST OF FIGURES xiii

CHAPTER I INTRODUCTION 2

1.1. Aims and Objectives 4

1.2. The Scope of Work 4

1.3. Research Funding 4

1.4. Publication 5

CHAPTER II BACKGROUND 6

2.1. Endoscopy 6

2.2. Semantic segmentation 8

2.2.1. Convolution Transpose Block 10

2.2.2. Linear Resize Upsampling Block 10

2.3. Data Augmentation 11

2.4. Parameter tuning for Supervised learning 11

2.4.1. Loss function 12

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

2.4.2. Optimiser 12

2.5. Evaluation Metrics 13

CHAPTER III RELATED WORKS 15

3.1. Relevant neural networks 15

3.1.1. BiSeNet, 2018 15

3.1.2. BiSeNetV2, 2020 16

3.1.3. STDC2-Seg, 2021 17

3.1.4. DDRNet-23-slim, 2021 18

3.1.5. BlazeNeo, 2022 19

3.2. Deep neural networks for GIM classification and segmentation 20

3.2.1. A multi-feature fusion method, 2021 22

3.2.2. Jigsaw augmentation for classification, 2022 23

3.2.3. Edge-guided path, 2021 24

3.3. Conclusion 25

CHAPTER IV CONCEPT AND RESEARCH METHODOLOGY 26

4.1. Data preparation 26

4.1.1. Location-wise hard negative sampling 29

4.2. Image preprocessing 31

4.2.1. Jigsaw augmentation 31

4.2.2. Label smoothing 31

4.2.3. Label smoothing with edge cut 33

4.3. Model improvement 34

4.3.1. Selective model for GIM image segmentation 34

4.3.2. Auxiliary head and loss 34

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

4.3.3. Separated threshold adjustment 36

4.3.4. Assembling multiple models for NBI and WLE modes 36

4.4. Model evaluation 38

4.5. Model deployment 38

CHAPTER V EXPERIMENTS AND RESULTS 39

5.1. Ablation study for single fine-tuned models on overall data 40

5.1.1. Selective model for GIM image segmentation 40

5.1.2. location-wise hard negative sampling 41

5.1.3. Jigsaw augmentation 41

5.1.4. Label smoothing 42

5.1.4. Auxiliary head and loss 43

5.2. Separated fine-tuned models for NBI and WLE modes with a classifier 45

5.3. Comparison on the entire dataset 47

5.3.1. Quantitative comparison 47

5.3.2. Qualitative comparison 48

5.4. Quantitative comparison on Abnormal EGD 49

5.5. Error Analysis of the winner model with the best techniques 49

5.5.1. Error Analysis of False Negative segmentation 49

5.5.2. Error Analysis of False Positive segmentation. 57

5.6. model deployment 58

CHAPTER VI CONCLUSION 59

CHAPTER VII APPENDICES 60

7.1. A single model with a classifier 60

7.1. An experiment of label smoothing with an edge cut and without. 60

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

REFERENCES 62

VITA.. 67

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

LIST OF TABLES

Page

Table 1. The experimental dataset consists of GIM (positive) and non-GIM (negative) images, with the non-GIM images categorized into three subclasses: normal EGD, abnormal EGD (non-GIM), and noise [28]. 27

Table 2. The number of endoscopic images as classified by the lighting modes: NBI and WLE [28].......... 28

Table 3. The number of endoscopic images as classified by gastric anatomical positions [28]. 28

Table 4. The brief navigation of the experiments and their winning models or techniques as the sequence. 39

Table 5. Performance comparison of different model architectures on the testing set. The LW-HNS technique was applied. Boldface refers to the winner [28]. 40

Table 6. Performance comparison of different hard negative sampling strategies on the testing set. Boldface refers to the winner [28].......... 41

Table 7. Performance comparison of different jigsaw augmentation strategies on the testing set. Boldface refers to the winner [28]. 42

Table 8. An effect of our label smoothing (gaussian edge smoothing) on the testing set. Boldface refers to the winner [28]. 42

Table 9. The effect of adding auxiliary head and loss into the network along with threshold adjustment on the testing set. There are three parts in the Table 8: (1) overall images, (2) WLE images, and (3) NBI images. Boldface refers to the winner [28]. 44

Table 10. Performance of NBI&WLE classification model using MobileNetV3s and speed in milliseconds (ms). 45

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Table 11. Performance comparison between a single BiSeNet model with an auxiliary head (separate thresholds), and separated fine-tuned BiSeNet models for NBI and WLE modes plus the classification model (MobileNetV3s). There are three parts in the Table 10: (1) overall images, (2) WLE images, and (3) NBI images. Boldface refers to the winner. 46

Table 12. Performance comparison between a single STDC2-Seg50 model with NBI&WLE classifier (MobileNetV3s and separate thresholds) and separated fine-tuned STDC2-Seg50 models for NBI&WLE modes plus the classifier (MobileNetV3s). There are three parts in the Table 10: (1) overall images, (2) WLE images, and (3) NBI images. Boldface refers to the winner. 47

Table 13. Quantitative comparison of BiSeNet model showing each stage of improvements: location-wise hard negative sampling (LW-HNS), the full jigsaw augmentation as (FJ), label smoothing (LS), and auxiliary head (AuxHead). Our model is also compared with the baseline. Boldface refers to the winner [28]. 50

Table 14. Quantitative comparison of STDC2-Seg50 model showing each stage of improvements: location-wise hard negative sampling (LW-HNS), the full jigsaw augmentation as (FJ), label smoothing with edge cut (LS), auxiliary head (AuxHead), and NBI&WLE classifier (MobileNetV3s). Our model is also compared with the baseline. Boldface refers to the winner. 51

Table 15. Quantitative comparison of BlazeNeo (DHA) model showing each stage of improvements: location-wise hard negative sampling (LW-HNS), the full jigsaw augmentation as (FJ), label smoothing (LS), and NBI&WLE classification model (MobileNetV3s). Our model is also compared with the baseline. Boldface refers to the winner [28]. 52

Table 16. Performance comparison of 133 abnormal EGD cases. Boldface refers to the winner from Table 13, 14, and 15. 55

Table 17. The effect of different label smoothing on the testing set was evaluated using STDC2-Seg50 model as the backbone. Boldface refers to the winner. 60

Table 18. Performance comparison between a single model with an auxiliary head and a single model with an NBI&WLE classifier (MobileNetV3s). There are three parts in the Table 12: (1) BiSeNet model, (2) STDC2-Seg50 model, and (3) BlazeNeo model. Boldface refers to the winner for each model.. 61

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

LIST OF FIGURES

Page

Figure 1. Upper gastrointestinal endoscopy [29]. 6

Figure 2. Absorption and scattering in tissue [30]. 7

Figure 3. The membrane of the human tongue. (A) White light image. (B) Narrow 8

Figure 4. The semantic segmentation of GIM. (A) Original image. (B) Prediction: white is GIM, and black is healthy. 8

Figure 5. U-net architecture [32]. 9

Figure 6. convolution transpose 2D with 1 stride and zero padding. 10

Figure 7. example of data augmentation applied on a GIM image. 11

Figure 8. Bilateral Segmentation Network; BiSeNet (a) Network Architecture (b) Attention Refinement Module (ARM). (c) Feature Fusion Module (FFM). 16

Figure 9. BiSeNetV2. The network has three main components: the two-pathway backbone, the aggregation layer, and the booster component [36]. 16

Figure 10. (a) Stem Block, (b) context embedding (c) aggregation layer [36]. 17

Figure 11. (a) STDC Segmentation network, The procedure in the dashed red box is STDC network. The procedure in the dashed blue box is Detail Aggregation Module. [38]. 18

Figure 12. DDRNets, "RB" denotes sequential residual basic blocks, "RBB" denotes the single residual bottleneck block, "DAPPM" denotes the Deep Aggregation Pyramid Pooling Module [39]. 19

Figure 13. (a) the detail of bilateral fusion, (b) the detail of DAPPM [39]. 19

Figure 14. Feature Aggregation: (a) Long skip connection (LSC), (b) Iterative Deep Aggregation (IDA), (c) Dense Iterative Aggregation (DIA), (d) Dense Hierarchical Aggregation (DHA). (e) Multi-headed BlazeNeo; (f) Receptive Field Block (RFB) [40]. 20

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Figure 15. multi-feature fusion model. AFM: attention feature module, FC: fully connected layer; BN: batch normalization [16]. 22

Figure 16. NBI images: (a) Non-GIM (b) GIM. The data is used by H. Li [16]. 23

Figure 17. data augmentation methods: Full Jigsaw and Partial Jigsaw. Partial jigsaw keeps the contents of the GT annotation [15]. 23

Figure 18. the result of Sobel's edge filter (a) raw image (b) image with Sobel's filter, the cropped images representing: (c) GIM (d) both GIM (the green arrow) and non-GIM (the red arrow) (e) non-GIM [26]. 24

Figure 19. Fast-SCNN modification with edge-guided; (a) core Fast-SCNN model (b) edge-guided block [26]. 24

Figure 20. The sample of (a) GIM image and (b) ground truth. 26

Figure 21. The experimental dataset is comprised of both GIM (positive) and non-GIM (negative) images, with the non-GIM images further classified into three subclasses: normal EGD, abnormal EGD (non-GIM), and noise [28]. 27

Figure 22. The process of Location-wise hard negative sampling (LW-HNS) comprises three primary steps. Firstly, an initial model is trained using the initial dataset. Secondly, it calculates the loss for negative cases that are not utilized (abnormal EGD and noise images). Lastly, the negative samples are chosen based on their losses in descending order for each category of gastric anatomical positions [28]. 29

Figure 23. The full jigsaw and the partial jigsaw augmentation as applied to GIM images [28]. 30

Figure 24. To implement the gaussian edge softening process, first a gaussian vector is created using the gaussian equation. Next, this gaussian vector is used to form the gaussian kernel by taking its outer product with another identical gaussian vector. Finally, this gaussian kernel is applied to the label images to soften their edges and reduce any noisy labels present [28]. 32

Figure 25. The first, second, and third steps of label smoothing with edge cut remain the same as the previous version. For the fourth, the label with edges softening is multiplied by the original ground truth to limit the non-zero label to being only inside the positive area. 33

Figure 26. The cross-section reveals the consequence of involving different label smoothing and the ground truth: (a) ground truth shows zero and positive regions with an edge at 250, (b) regular label smoothing displays over-labeling on the negative side, (c) label smoothing with edge cut preserves the negative site. 33

Figure 27. The enhanced model incorporated an auxiliary head for classifying imaging modes and auxiliary training. In (1), there is the auxiliary head. In (2), decision thresholds can be altered independently for NBI and WLE modes [28]. 35

Figure 28. The models that supported NBI and WLE modes are proposed in 3 approaches: (1) The single model with the auxiliary head, (2) The single model with the lightweight classification model, and (3) the separated fine-tuned models for NBI and WLE modes with the lightweight classification model. 37

Figure 29. The instances show non-GIM images that are incorrectly identified as false positives due to structural bias. GIM is likely to appear in some particular areas (structures), such as the hole of gastric autumn, so the prediction is biased by the hole or other gastric structures, regardless of the GIM texture [28]. 41

Figure 30. The distribution of GIM confidence per image (taking an average for all pixels) on the validation set: (a) all images contain imaging modes and (b) separated imaging modes. Since NBI is more confident than WLE, it is preferable to use separate thresholds in this case [28]. 43

Figure 31. The output of the WLE fine-tuned model is overshot compared to the ground truth, yielding a lower IoU score relative to the universal model with an auxiliary head. 46

Figure 32. On the test dataset, a qualitative comparison of the proposed method for BiSeNet (ResNet-18) is illustrated [28]: (a) original image; (b) ground truth; (c) BiSeNet; (d) BiSeNet+LW-HNS; (e) BiSeNet+LW-HNS+FJ; and (g) BiSeNet+LW-HNS+FJ+LS+AuxHead. There are 12 instances (rows) of NBI GIM (1st–4th), WLE GIM

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

(5th–8th), NBI non–GIM with abnormal (9th–10th), and WLE non–GIM with abnormal (11th–12th). (c) cite the preceding work [28]. 53

Figure 33. Qualitative comparison of the winner method for each model and the original baseline model on the test dataset: (a) original image, (b) ground truth, (c) BiSeNet+LW-HNS, (d) BiSeNet+LW-HNS+FJ+LS+AuxHead, (e) STDC2-Seg50+LW-HNS, (f) STDC2-Seg50+LW-HNS+FJ, (g) BlazeNeo+LW-HNS, (h) BlazeNeo+LW-HNS+FJ+LS +MobileNetV3s. There are 12 examples (rows) composed of NBI GIM (1st-4th), WLE GIM (5th-8th), NBI non-GIM with abnormal (9th-10th), and WLE non-GIM with abnormal (11th-12th). 54

Figure 34. The false negative images of GIM in NBI mode: 3 images of incisura, 2 images of corpus, and 1 image of cardia and fundus. 56

Figure 35. The false negative images of GIM in WLE mode: 3 images of incisura, 3 images of corpus, and 2 images of antrum. Some images can be overlooked easily. 56

Figure 36. The false positive images of GIM in NBI mode: 2 images of abnormal EGD and one of normal EGD. 57

Figure 37. The false positive images of GIM in WLE mode: 3 images of abnormal EGD, 3 images of normal EGD, and one noise. 57

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

CHAPTER I
INTRODUCTION

Gastric cancer is among the most prevalent forms of cancer. According to the global incidence of gastric cancer in 2020, more than a million cases were recorded, and approximately 700,000 individuals passed away [1]. However, a delayed cancer diagnosis can result in delayed treatment, thereby increasing mortality risk [2]. Gastric cancer must be diagnosed as soon as possible. With early detection, treatment is more effective, and the 5-year survival rate is expanded to nearly 95% [3].

The detection of gastrointestinal metaplasia (GIM) is a procedure performed for the early detection of gastric cancer. The evolution of GIM begins with secretory cells that produce hydrochloric acid in oxyntic mucosa, causing a mutation in the intestinal mucosa [4]. GIM appears to be easily ignored due to the morphology of flat mucosa and few differences from normal mucosa, particularly during esophagogastroduodenoscopy (EGD) with white light endoscopy (WLE) [5]. Narrow-band imaging (NBI) was developed as an image enhancement technique using specific forms of narrow-band light in order to highlight the microvascular structures on the mucosal surface [6]. The addition of NBI technology has increased the detection sensitivity of GIM by 37% (from 53% to 90%) [7]. Manual GIM detection significantly depends on clinical experience. Even with image enhancement, interobserver variability is still substantial [8].

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

In medical image processing, the automatic diagnosis of abnormal EGD has become widespread in recent years. In the past few decades, handcrafted feature-based detection approaches [9, 10]. T. Kanesaka [9] created a computer-aided diagnosis (CADx) of early gastric cancer with the use of a support vector machine to identify early gastric cancer in M-NBI mode. The features were generated by partitioning grayscale images, selecting eight grey-level cooccurrence matrix (GLCM) features [11], and then calculating the variance coefficient for eight GLCM features. The sensitivity and precision of CADx were 96.7% and 98.3%, respectively. Subsequently, a deep learning (DL) approach emerged in the field of medicine, particularly in the gastrointestinal tract (GI). Thus, DL became the primary method for research into the automatic diagnosis of endoscopy images. The detection of polyps is one of the most well-known research topics in lower GI [12]; in upper GI, the system of early gastric neoplasms was developed [13, 14].

All previous works on GIM classification [15-21] and segmentation [22-25] were unable to attain real-time inference speed. In practical use, such studies could not be implemented. Recent investigations [26, 27] have demonstrated that inference speed and accuracy can be achieved. Nevertheless, it is apparent that performance could be further enhanced by addressing additional issues found in real-world use cases, such as different colour modes (NBI and WLE), unrelated events (bubbles, tools, etc.), more abnormal EGD (ulcer and erosion), and noise labels.

The ultimate goal is to enhance the performance of GIM semantic segmentation using the real-time system in a more practical approach by considering bias (structural bias and location bias), noise (noisy labels), and other abnormal EGD, resulting in more excellent practicality and generalisation of our model. The further contributions are as follows:(1) Based on inference speed and segmentation performance, a comprehensive selection of segmentation models is undertaken in order to select the most optimal model. (2) The model is modified by assembling an auxiliary head to support multiple imaging modes; additionally, an additional loss is added to guide the model regarding colour modes. (3) The "location-wise hard negative sampling" method is used to reduce a location bias in gastric morphology. (4) A jigsaw augmentation is used to reduce structural bias, enabling the model to establish a direct correlation with the GIM feature. Occasionally, the model makes a prediction using a stomach feature but not GIM morphology. (5) GIM's noisy label is reduced by label smoothing, also known as gaussian label edge softening, in an effort to enhance performance for ambiguous ground truth. (6) Lastly, by installing the auxiliary head of the imaging modes classification to our model, performance can be enhanced by different decision thresholds for each imaging mode.

1.1. Aims and Objectives

1. To propose a deep learning network for real-time GIM semantic segmentation that can effectively handle multiple abnormalities: location bias, structural bias, noisy labels, etc., in order to assist endoscopists.
2. To evaluate the performance of the proposed network and techniques for GIM segmentation on challenging scenarios: multiple colour modes, multiple lesions, and noise.
3. To evaluate the inference speed of the proposed network, which must exceed 25 fps.

1.2. The Scope of Work

1. Evaluate the proposed deep learning network along with the following.
    a. Experiment on our private dataset of GIM obtained from the Center of Excellence for Innovation and Endoscopy in Gastrointestinal Oncology, Chulalongkorn University, Thailand.
    b. GIM images in our dataset were acquired from 136 patients by expert endoscopists.
    c. Abnormal EGD is only included GIM, erosion and ulcer; multiple imaging modes are only WLE and NBI.
2. The proposed deep learning network can differentiate between GIM region and healthy background that solely applies to EGD.
3. The promised inference speed of the real-time GIM segmentation is 25 FPS.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

1.3. Research Funding

This research project was funded by the National Research Council of Thailand (NRCT; N42A640330), Chulalongkorn University (CU-GRS-64), and Chulalongkorn University (CU-GRS-62-02-30-01) and supported by the Center of Excellence in Gastrointestinal Oncology, Chulalongkorn University annual grant. It was also funded by the University Technology Center (UTC) at Chulalongkorn University.

1.4. Publication

- Pornvoraphat, P., et al., *Real-time gastric intestinal metaplasia diagnosis tailored for bias and noisy-labeled data with multiple endoscopic imaging.* Computers in Biology and Medicine, 2023. **154**: p. 106582. [28]
  - ○ Computers in Biology and Medicine Journal, Elsevier, Q1.
  - ○ Impact Factor = 6.698.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

CHAPTER II

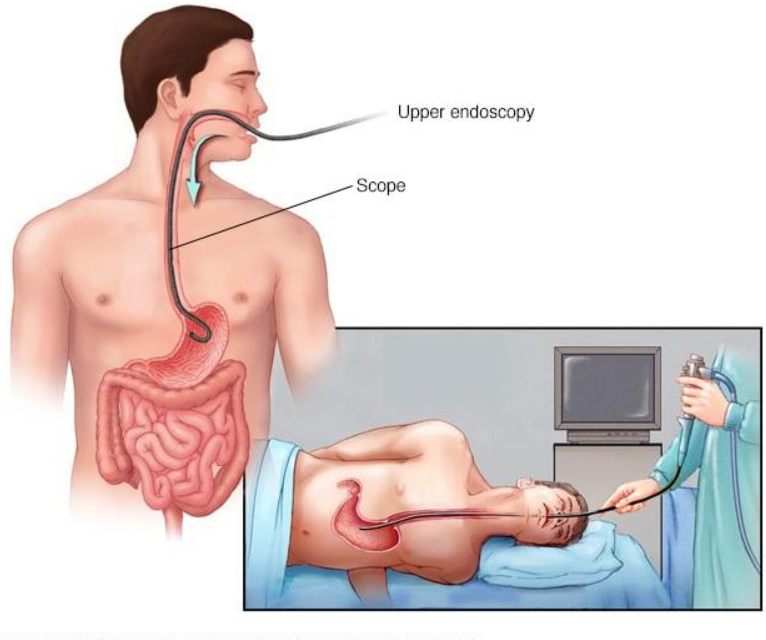
BACKGROUND

In this chapter, the background knowledge related to the thesis is presented. Endoscopy, semantic segmentation, data augmentation, parameter optimization, and evaluation for semantic segmentation and classification are explained.

2.1. Endoscopy

Gastrointestinal (GI) endoscopy is a medical procedure that allows endoscopists to assess, diagnose, and treat GI illnesses through real-time images. GI endoscopy can be specified as upper GI and lower GI. Examples of upper GI are esophagus, stomach, duodenum, and jejunum. For lower GI, it is included rectum, colon, and terminal ileum. Typically, the terms that are relevant to endoscopy are as follows:

- Colonoscopy is related to an examination of the lower GI.
- Esophagogastroduodenoscopy (EGD) is associated with an investigation of the upper GI.



© MAYO FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH. ALL RIGHTS RESERVED.

Figure 1. *Upper gastrointestinal endoscopy [29].*

To observe the upper digestive tract, endoscopists use a tiny camera called an endoscope to insert through the throat into the esophagus and stomach, see Figure 1. At the end of the endoscope, the light source illuminates the digestive tract, enabling specialists to conduct the necessary procedures. The visual signal is transmitted via a flexible tube and shown on display. Generally, The light source for EDG can be categorised into two groups: white light endoscopy (WLE) and narrow-band imaging (NBI) [30].

WLE is a technique that uses normal xenon light as a light source. In contrast, NBI applies an NBI filter at the light source to limited wavelengths (allow 415 and 540 mm in the centre wavelength narrow band). In biology, the different wavelengths have different behaviours in biological tissue, see Figure 2. When red light enters the tissue, it diffuses widely and deeply. Meanwhile, the blue light is the opposite. The colour of the gastrointestinal mucosa is mainly defined by haemoglobin; haemoglobin is a type of chromophore mainly found in the blood. With the wavelengths of NBI that correspond to the haemoglobin absorption maxima, the capillary networks are revealed more clearly and effectively (see Figure 3).

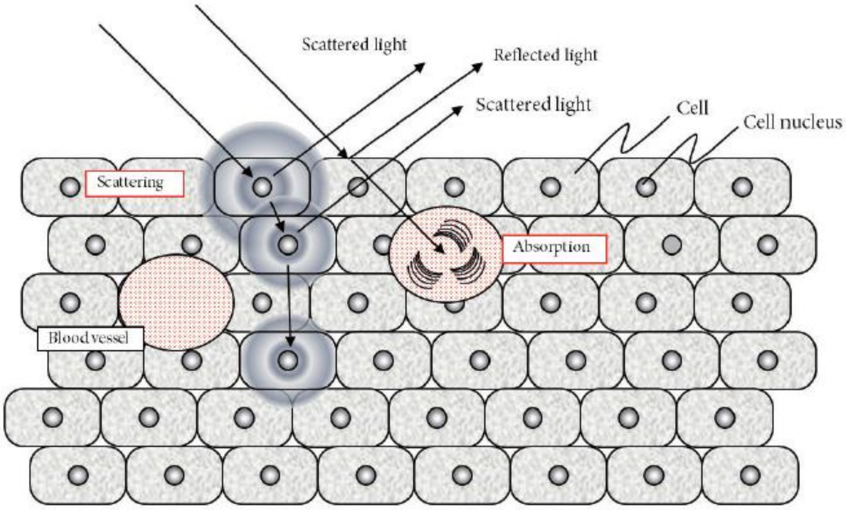


Figure 2. *Absorption and scattering in tissue [30].*

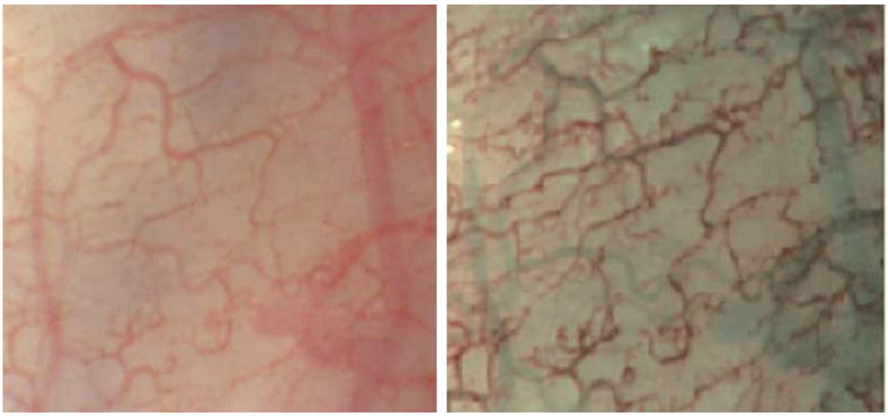


__Figure 3.__ The membrane of the human tongue. (A) White light image. (B) Narrow band imaging [30].

2.2. Semantic segmentation

The objective of semantic segmentation is to assign each pixel in an image to a given category by labeling each pixel in the same shade for an identical class, see Figure 4. For each photo, a deep neural network (DNN) is deployed to transform the original image into a semantic label image with the exact dimensions as the input. However, the number of output channels is equal to the number of classes. The network relies on using downsampling and upsampling to generate a semantic output. Here, U-Net [31] is a suitable representation of segmentation architecture for semantic tasks.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

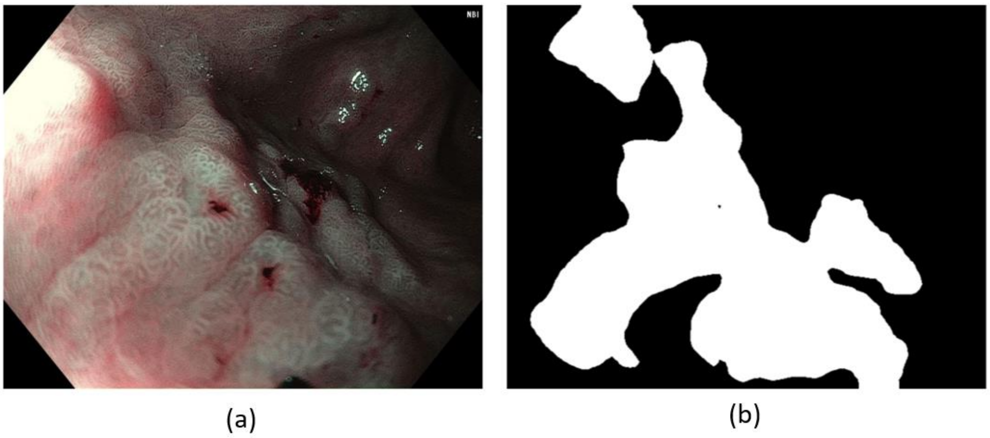


__Figure 4.__ The semantic segmentation of GIM. (A) Original image. (B) Prediction: white is GIM, and black is healthy.

U-Net is one of the most primal architectures for semantic segmentation. As seen in Figure 5, the model comprises a downsampling half and an upsampling half. The downsampling haft (encoder) resembles the conventional design of a convolutional neural network (CNN), which is often used for classification applications. The functionality of the encoder is to extract high-level features: context information. Together, upsampling (decoder) and skip connection permit the model to construct a segmentation map from each level of features and contextual information.

Downsampling architecture (encoder) is adopted from VGGNet [30]. For each block of convolution, ReLU, and max-pooling layers, the output resolution is dropped by one-half while channels are doubled. Each pyramid output of the encoder is utilised by concatenating it with decoder outputs for each equivalent level. All of this enables the encoder to summarise the image's context information, which is essential for semantic segmentation.

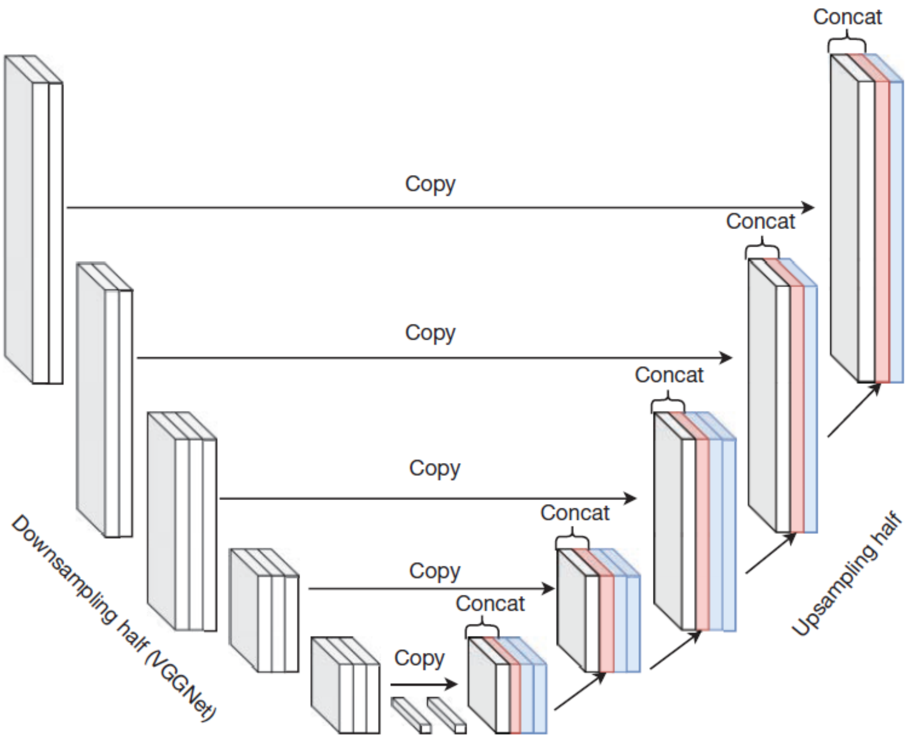


Figure 5. U-net architecture [32].

Due to the fact that context information is a summarization of a picture that is deficient in spatial detail, half-upsampling is utilized to restore such loss of detail. For each upsampling block, the concept is to upsample the prior context data received from the previous block and then concatenate it with the pyramid feature from the encoding at the same level to recover the detail. Two approaches exist for the upsampling block:

2.2.1. Convolution Transpose Block

Convolution Transpose Block is composed sequentially of convolution transpose 2D, concatenation, and double convolution block. The double convolution consists of repeated convolution layers, batch normalisation, and ReLU layers. As demonstrated in Figure 6, the convolution transpose 2D leverages a kernel to multiply the scalar value of the image pixel and clones it to the region corresponding to the pixel location. The product results are aggregated to get the final outcome.

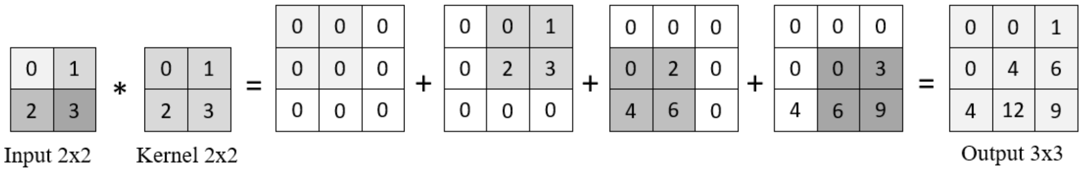


Figure 6. *convolution transpose 2D with 1 stride and zero padding.*

CHULALONGKORN UNIVERSITY

2.2.2. Linear Resize Upsampling Block

In spite of adopting convolution transpose as the core, resize upsampling can provide an alternate outcome. After linear upsampling is applied, concatenation and double convolution are performed.

2.3. Data Augmentation

Data augmentation is utilized to expand the quantity of data by introducing slightly modified duplicates or synthetic data. The primary objective is to increase regularisation and reduce overfitting when training with data. Data augmentation can also determine as an oversampling technique in data science. This augmentation is specified into two groups: weak augmentation (shifting, scaling, rotation, flipping, and transposing) and strong augmentation (adding noise, distortion, sharpening, and blurring).

In the studies, each augmentation (Figure 7) is carefully utilised and adjusted in its hyperparameter; thus, all augmentation must be compatible and reasonable with EGD. For instance, when performing colour adjustment, the tone must be in a range of ordinary endoscopic images for a particular light mode. The blue and green tones are not accepted. For this thesis, Albumentations library is adopted to perform data augmentation.

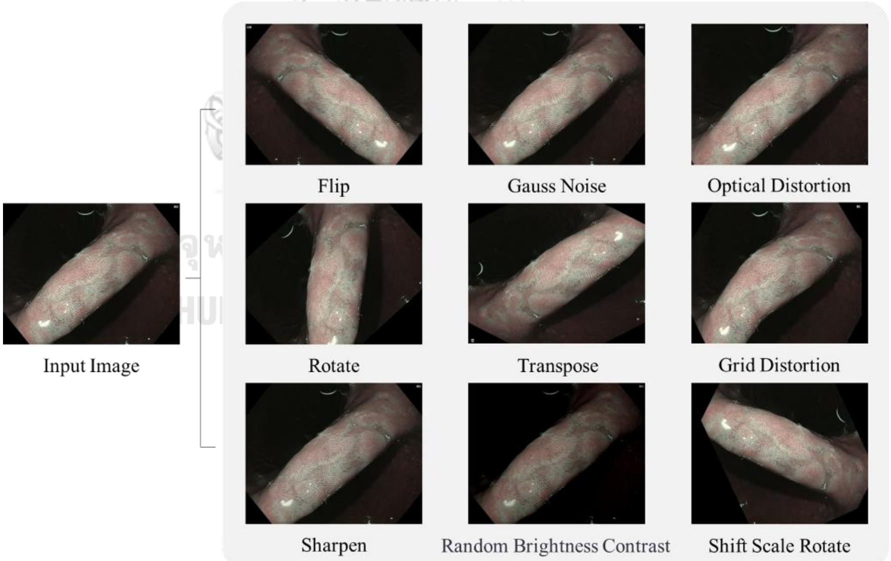


Figure 7. *example of data augmentation applied on a GIM image.*

2.4. Parameter tuning for Supervised learning

Parameter tuning for supervised learning is a learning process of models (deep neural network, traditional machine learning) by learning from given pairs of input and label output. The learning process is various and depends on the type of the model. In DNN, the training process uses a loss function to minimise an error in the model by gradually adjusting its weight. The weight adjustment depends on the selected optimiser; most computations require gradients for each layer. The gradients can be estimated using a backpropagation algorithm since the model is too complicated to compute directly.

2.4.1. Loss function

The loss function for supervised learning can be categorised into two major groups: classification and regression. For semantic segmentation tasks, it shares the same loss as classification; thus, this study gives attention. Here binary and category cross-entropy loss can be deployed as the main loss:

$$\mathrm{L} = -\sum_{n\in N}\sum_{c\in C} w_c log \frac{\exp(x_{n,c})}{\sum_{i\in C}\exp(x_{n,i})} y_{n,c} \tag{1}$$

The loss takes pairs of labels $y_{n,c}$ and model prediction $x_{n,c}$ for its computation. w is class weight, C is all classes, N can be all samples in minibatch or all pixels of images. Auxiliary loss can also add to the objective function for specific purposes: regularisation, training assisting etc.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

2.4.2. Optimiser

In DL, the optimiser functionality is to update weight progressively until the loss function is converted; it reaches the local minimum. For this process, the gradients are required for updating. The final result is not deterministic depending on hyperparameter settings: initial weights, learning rate etc. The most well-known optimiser for DL is stochastic gradient descent (SGD). SGD implements randomly selected batches of data to update weight. The weight optimisation is described as follows:

$$\mathrm{w} = \mathrm{w} - \eta\nabla L(w) \tag{2}$$

Where w is model weight, η is learning rate, ∇ is gradient and L is loss function.

2.5. Evaluation Metrics

$\mathrm{Io}U_{image}$ was used to evaluate the segmentation of a single image by computing intersection over union pixels of ground truth and prediction of GIM segmentation. In Equation (3), the $\mathrm{Io}U_{image}$ can be defined as follows:

$$\mathrm{Io}U_{image} = \frac{\sum_{i\in I} g_i \cdot p_i}{\sum_{i\in I} g_i + \sum_{i\in I} p_i - \sum_{i\in I} g_i \cdot p_i} \quad (3)$$

where i represents a specific pixel of an image. g_i and p_i correspond to the i-th pixel on a ground truth and a prediction, respectively. The value of g_i and p_i for each i-th pixel is either 1 (GIM) or 0 (non-GIM). For each GIM image, the $\mathbf{IoU_{image}}$ is averaged to represent the overall segmentation performance of GIM and is denoted as IoU. In the performance of the models in negative images, error_{image} is used to distinguish between TN and FP on an image using GIM prediction pixels over all image pixels. In Equation (4), error_{image} is defined. A mean of error_{image} is provided to demonstrate the model's confusion of negative prediction and denoted as “Error:”

$$\mathrm{error}_{image} = \frac{\sum_{i\in I} p_i}{\sum_{i\in I} 1} \quad (4)$$

where $i \in I$ represents an individual pixel of an image and $p_i = 1$ is GIM, the segmentation efficacy of each image determines the $p_i \in \{0,1\}$ GIM detection. GIM images are considered true positive (TP) if $\mathrm{Io}U_{image}$ is greater than 30%; otherwise, the outcome is false negative (FN). True negative (TN) corresponds to non-GIM images with a “Error” percentage of less than 1%. False positives (FP) are referred to non-GIM images in which the excess zone of GIM prediction error_{image} is greater than 1% of the entire image. GIM detection evaluation metrics include accuracy, specificity, sensitivity, negative predictive value (NPV), positive predictive value (PPV), and F1-score (only for GIM). The evaluation metrics are represented as follows:

$$\mathrm{accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

$$\text{recall}_{GIM}(sensitivity) = \frac{TP}{TP+FN} \tag{6}$$

$$\text{recall}_{non-GIM}(specificity) = \frac{TN}{TN+FP} \tag{7}$$

$$\text{precision}_{GIM}(PPV) = \frac{TP}{TP+FP} \tag{8}$$

$$\text{precision}_{non-GIM}(NPV) = \frac{TN}{TN+FN} \tag{9}$$

$$\text{F1} = \frac{2 \times recall \times precision}{recall + precision} \tag{10}$$

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

CHAPTER III

RELATED WORKS

This chapter presents the deep neural networks relevant to the thesis. In medical image analysis, the most popular application of deep neural networks related to gastric intestinal metaplasia (GIM) can be divided into two categories: image classification and semantic segmentation. Here the relevant neural networks and techniques used for GIM classification and segmentation are reviewed.

3.1. Relevant neural networks

This section presents the relevant neural networks: BiSeNet, BiSeNetV2, STDC1-Seg, and DDRNet-23-slim. All of these models accomplish incredible inference speed and performance on cityscape datasets viz. achieve real-time segmentation. Plus, BlazeNeo also attains excellent efficiency on its own polyps dataset.

3.1.1. BiSeNet, 2018

Figure 8 illustrates the Bilateral Segmentation Network (BiSeNet) architecture [33], which consists of a spatial path, context path, and feature fusion module. Simultaneous extraction of both spatial and context information makes model inference fast. The spatial path was designed to extract spatial information and retain input resolution, enabling a highly detailed segmentation output. In contrast, the context path can extract high-level features and aggregate content data in an image. The spatial path is a shallow module with three convolution layers, batch normalisation, and ReLU. The context path adopts a lightweight model as its backbone, of which there are two variants: Xception [34] and ResNet-18 [35]. The backbone can provide two resolution outputs: 16x and 32x downsampling features that are aggregated into the context output using a U-shaped design. Last but not least, the feature fusion module (FFM) can incorporate spatial and content features into the model's prediction.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

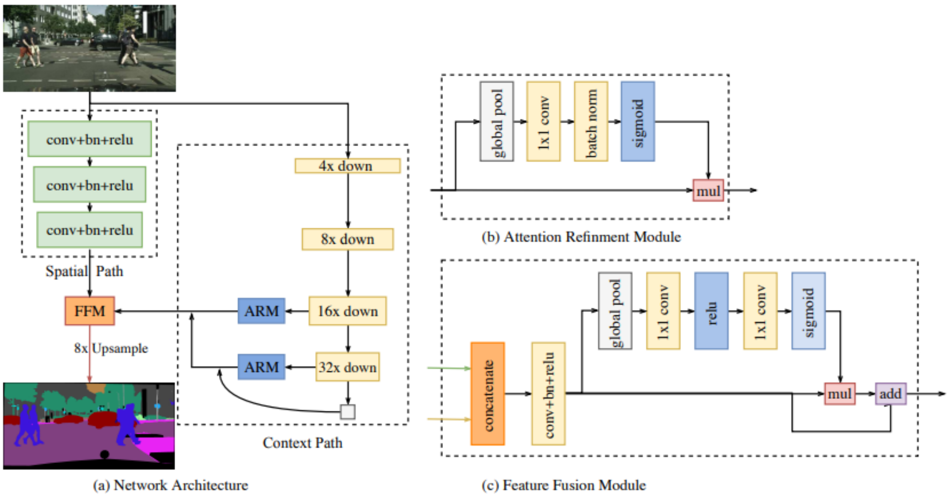


Figure 8. *Bilateral Segmentation Network; BiSeNet (a) Network Architecture (b) Attention Refinement Module (ARM). (c) Feature Fusion Module (FFM).*

3.1.2. BiSeNetV2, 2020

BiSeNetV2 [36] employs the same bilateral segmentation architecture as BiSeNet (Figure 9), which is composed of a spatial path, context path, and feature fusion module. The difference is those three elements were upgraded. The spatial path uses VGGNet without skip connection instances of the three blocks of convolution, bath normalization and ReLU layers that were proposed in BiSeNet. The context path was redesigned to be a lighter-weight structure. Rather than using the light-weight backbone: Xception and MobileNet [37].

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

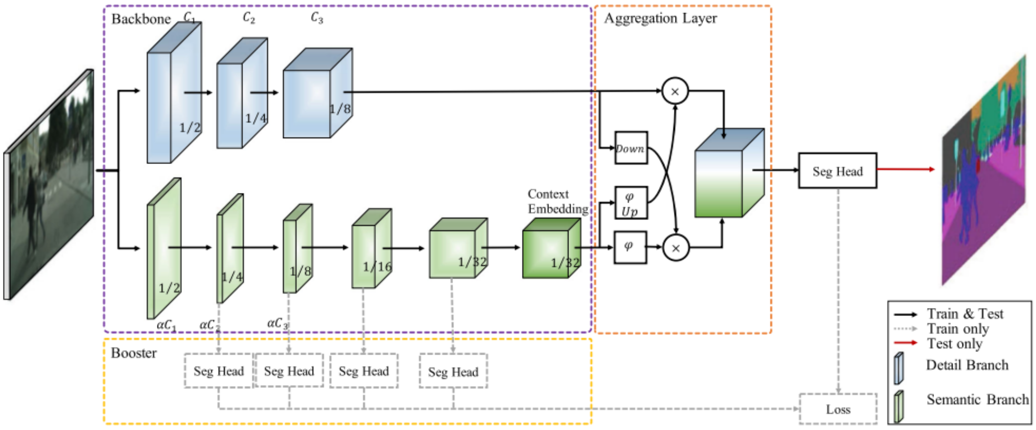
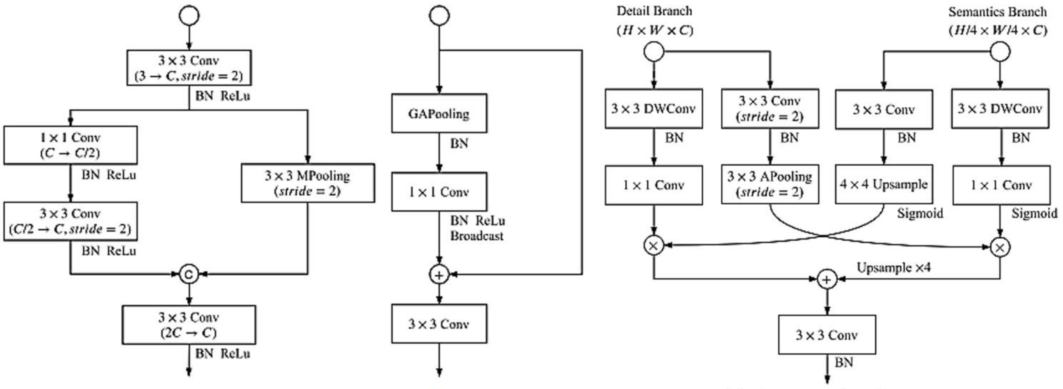


Figure 9. *BiSeNetV2. The network has three main components: the two-pathway backbone, the aggregation layer, and the booster component [36].*

Stem Block was adopted for each block of an encoder, see Figure 10a. Its functionality is to capture two different receptive fields at once and then concatenate both branches as the final output. This structure has efficient computation cost and practical feature representation. At the end of the new encoder, the new context embedding (CE) was proposed to embed the global contextual information; Figure 10b. Finally, the feature fusion of the previous one is replaced by a novel aggregation layer, see Figure 10c.



(a) Stem Block **(b)** CE Block **(c)** Aggregation Layer

Figure 10. *(a) Stem Block, (b) context embedding (c) aggregation layer [36].*

3.1.3. STDC2-Seg, 2021

STDC2-Seg [38] uses the same bilateral segmentation architecture as the two previous versions. Still, the backbone of the encoder was upgraded from the first version, and the spatial path is no longer, see Figure 11a. The brand-new backbone, STDC, was invented to replace the ResNet-18 of the first version. The accuracy of STDC1 and STDC2 outperform ResNet-18 on ImageNet by 4% and 6%, respectively. The spatial path is skipped and connected to FFM; thus, the new connection transfers detailed information directly. In order to accomplish this, detailed guidance is applied to guide the low-level layers of detailed information. Detailed guidance needs to have a detailed loss (Figure 11b) and detailed ground truth for optimizing using binary cross-entropy loss and dice loss Detailed ground truth can be constructed by applying Laplacian convolution for each stride (1,2,4), fusion with 1x1

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

convolution, and then adopting a threshold of 0.1 to convert the predicted to binary, see Figure 11c.

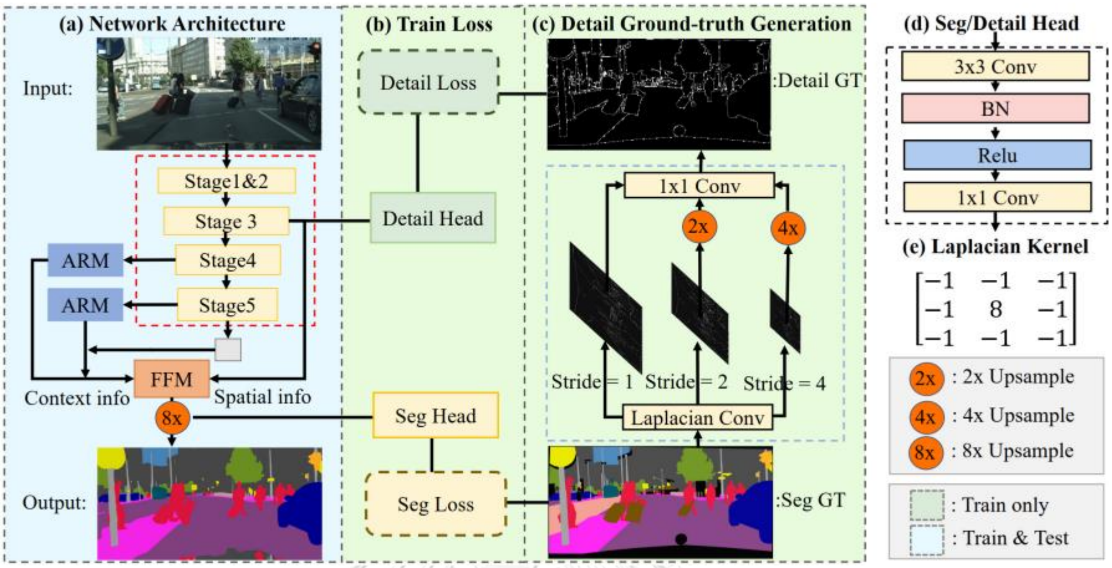


Figure 11. *(a) STDC Segmentation network, The procedure in the dashed red box is STDC network. The procedure in the dashed blue box is Detail Aggregation Module. [38].*

3.1.4. DDRNet-23-slim, 2021

Deep Dual-resolution Network (DDRNet) [39] has three components: high-resolution, low-resolution branches and deep aggregation pyramid pooling module (DAPPM), see Figure 12. The low-resolution branches adopted a modified ResNet as the backbone, replacing a 7×7 convolutional layer with two 3×3 convolutional layers. The high-resolution branch was added after the second basic residual block (RB) of the backbone. In the high-resolution branch, the image resolution is constant at 1/8 of the original size since all convolution layers have a stride of 1. During each RB, there has bilateral fusion (see Figure 13a) to cross-exchange information between them. The high resolution contains spatial features; meanwhile, the low resolution contains context features. At the end of low-resolution branches, DAPPM (Figure 13b) is utilized for performing multiple-receptive fields; thus, rich context information is gathered. The performance of DDRNet-23-slim on ImageNet exceeds BiSeNet (ResNet-18) by 2%.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

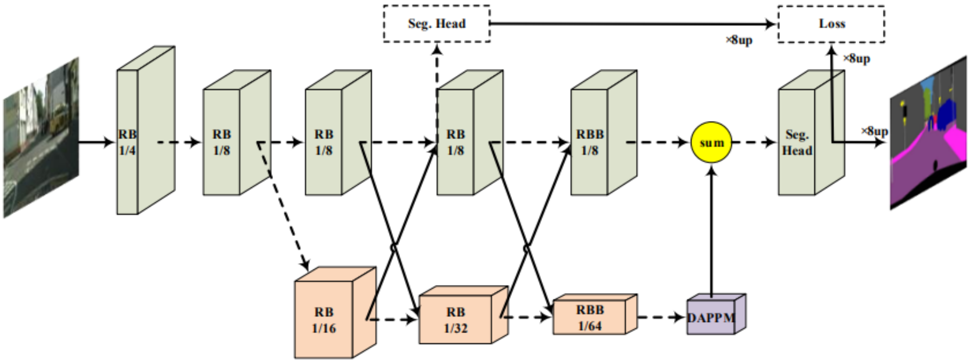
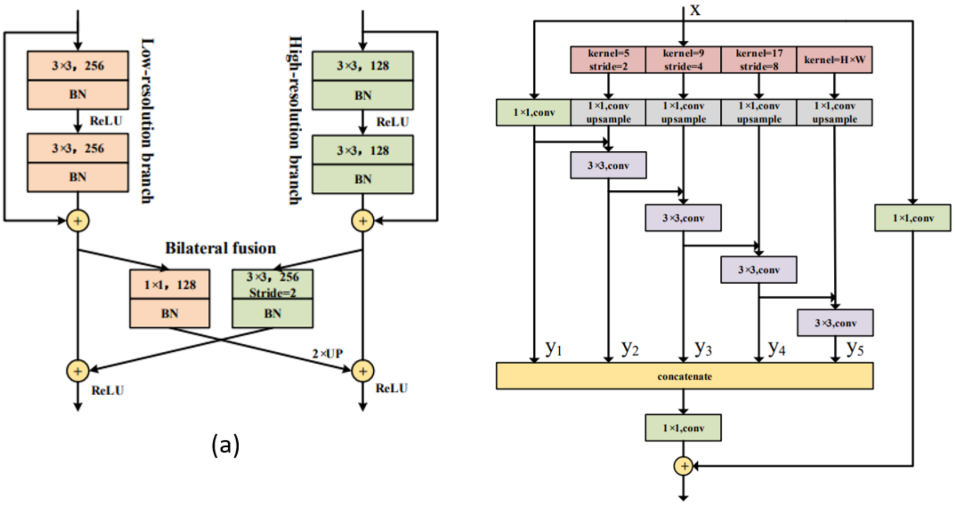


Figure 12. *DDRNets, "RB" denotes sequential residual basic blocks, "RBB" denotes the single residual bottleneck block, "DAPPM" denotes the Deep Aggregation Pyramid Pooling Module [39].*



(a)

(b)

Figure 13. *(a) the detail of bilateral fusion, (b) the detail of DAPPM [39].*

3.1.5. BlazeNeo, 2022

One of BlazeNeo's contributions is a new DNN for polyps segmentation with a light-weight encoder-decoder and feature aggregation, see Figure 14. For the light-weight encoder, BlazeNeo [40] adopted HarDNet-68 as a backbone. The performance on ImageNet outperforms ResNet-18 and VGG-16 by 6.6% and 2.8%. Testing on GTX1080 at 1024x1024, the GPU time of HarDNet-68 (32.6 ms) is impaired to ResNet-18 by 19 ms but enhanced from VGG-16 by 46 ms.

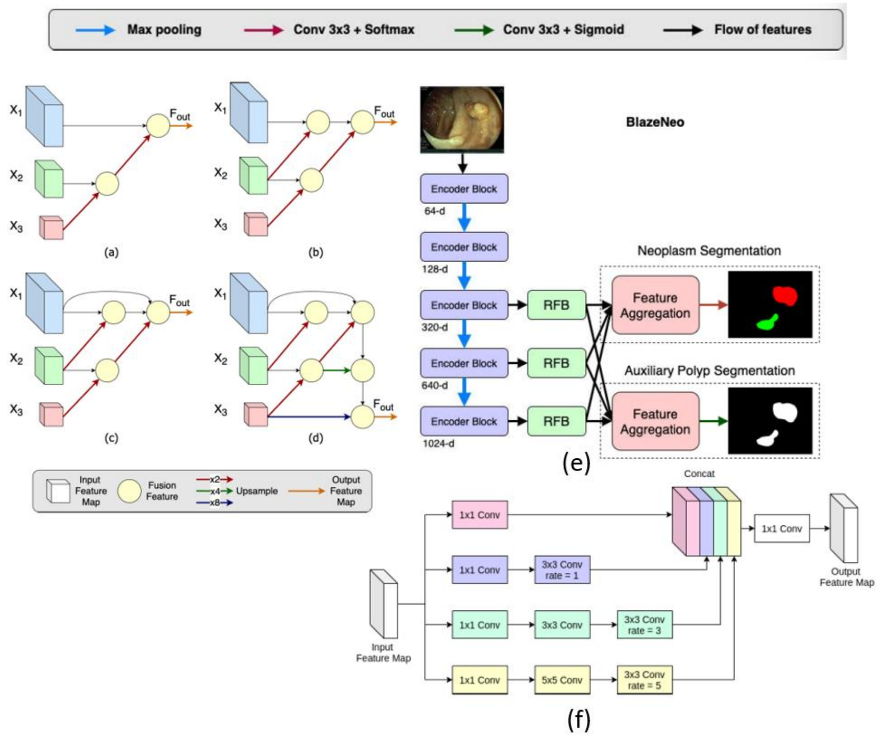


***Figure* 14.** *Feature Aggregation: (a) Long skip connection (LSC), (b) Iterative Deep Aggregation (IDA), (c) Dense Iterative Aggregation (DIA), (d) Dense Hierarchical Aggregation (DHA). (e) Multi-headed BlazeNeo; (f) Receptive Field Block (RFB) [40].*

After the encoder block of 310-d, 640-d, and 1024-d, the receptive field block (RFB) is employed. RFB uses multi-branch convolution to improve efficiency inspired by the human visual cortex see detail in figure 14 (f). Each output of RFB is accumulated by feature aggregation: LSC, IDA, DIA, and DHA. DHA showed the best performance among other approaches.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

3.2. Deep neural networks for GIM classification and segmentation

In this decade, the development of DL for medical image analysis has been most notable, resulting in a significant performance increase. Prior DL for GIM analysis primarily focused on image classification, with only a handful of systems capable of semantic segmentation. T. Yan [20] proposed a GIM diagnostic system based on NBI and M-NBI images in 2020. The system could classify GIM and non-GIM images with a sensitivity of 91.9%, a specificity of 86.00%, and an accuracy of 86.00%. A year later,

H. Li [16] used a multi-feature fusion method combining the features of RGB, Hue Saturation Value (HSV), and Local Binary Pattern (LBP) images to improve classification performance in NBI mode. N. Lin [17] was interested in conditioning a GIM classification model on white light endoscopic (WLE), which is typically more difficult than NBI mode, in the same year. Tao Yu introduced multi-label recognition of gastric lesions on WLE [15] in 2022. The DL model could categorise endoscopic images of the gastrointestinal tract into normal gastric mucosa, atrophic gastritis, GIM, and gastric malignancy. Another study [18] developed a classification model for GIM subtypes as healthy, mild, moderate, and severe. P. K. Wong [19] proposed a comprehensive learning system that aggregated five output classifiers into the system's final output to improve classifier performance.

As for classification tasks, GIM diagnosis is substantially more developed. GIM segmentation, in contrast, is somewhat outdated. Few investigations have been conducted regarding segmentation tasks. GIM segmentation on NBI images was first introduced in 2017 [22] using hue energy to represent global colour features and texture energy for describing local microvascular texture features to enhance segmentation performance. Two years later, C. Wang [23] proposed using the W-Deeplab model to segment GIM. W. Du [24] established the segmentation network ResUnet in 2021, replacing the VGG block of U-Net with ResBlock. The model was designed to segment early gastric cancer and achieved 92.22% IoU, while U-net achieved 91.45% IoU. K. Qiu [25] utilised an enhanced U-Net-based network for segmenting gastric precancerous lesions in ME-NBI mode, including inflammation, GIM, low-grade neoplasia, and early cancer the following year. The model achieved a 96% F1 score.

Even though the prior study has demonstrated exceptional segmentation performance, real-time systems have not yet been completely developed. Segmentation of GIM in real-time first appeared in [26, 27]. Consequently, the paper [26] introduced the Fast-SCNN network, which was explicitly designed for GIM using an edge-guided path to enhance performance. The accuracy of real-time semantic segmentation was improved by integrating four techniques [27].

The engineer's interest techniques from the prior review of GIM are listed, which is explained in more detail. (1) A multi-feature fusion method proposed for classifying GIM in NBI mode by H. Li [16], 2021. (2) The jigsaw augmentation for multiple-label classification of cancer-related lesions adopted by Tao Yu [15], 2022. (3) Edge-guided path for fast semantic segmentation of GIM proposed by V. Siripoppohn [26], 2021.

3.2.1. A multi-feature fusion method, 2021

Typically, endoscopists consider multiple features for diagnoses such as colour, texture and shape. Thus, the attention feature module (AFM) is proposed to combine the feature of RGB image, hue saturation value (HSV) image representing colour and local texture feature (LBP) image representing texture [41]. After combining, the rest of the network is classic fully connected layers, batch normalisation and ReLU, see Figure 15. The performance after utilising all features of RGB, HSV, and LBP outperforms original RGB images by 6.3% on the F1 score. Again, this implementation is based on NBI mode (Figure 16), and multiple lesions outside GIM are not included, which texture alone is straightforward.

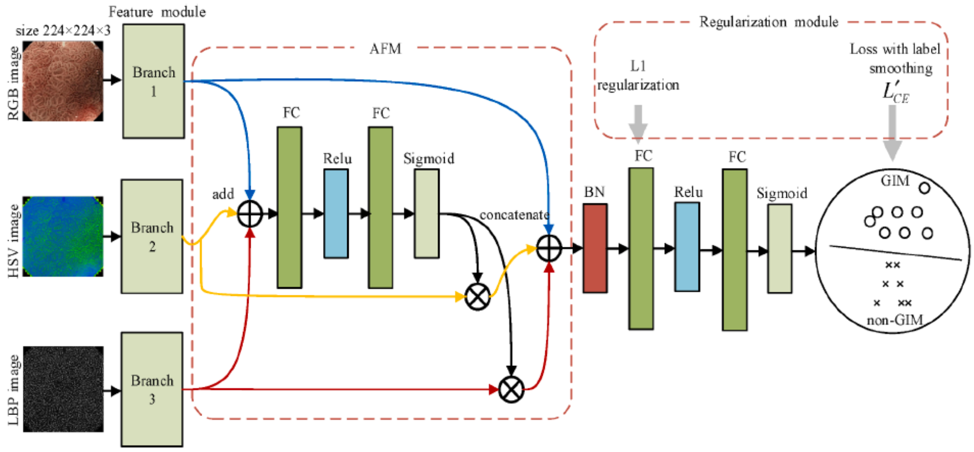


Figure 15. multi-feature fusion model. AFM: attention feature module, FC: fully connected layer; BN: batch normalization [16].

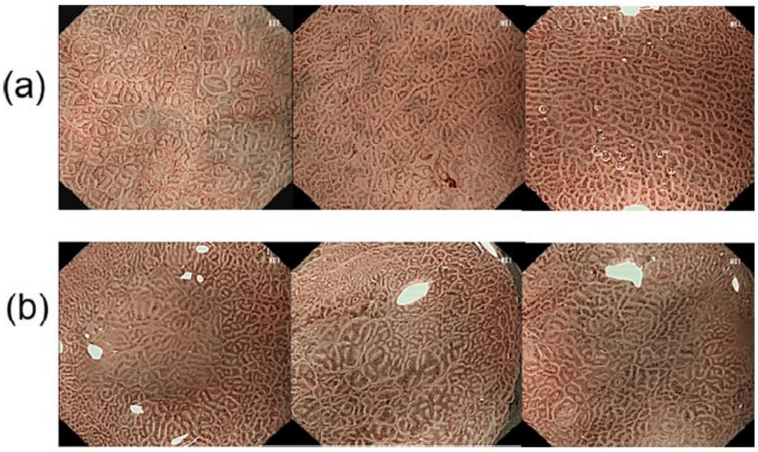


***Figure* 16.** *NBI images: (a) Non-GIM (b) GIM. The data is used by H. Li [16].*

3.2.2. Jigsaw augmentation for classification, 2022

Jigsaw augmentation, used for fine-grained image recognition, divides an image into many patches and then randomly reassembles them into a new image. This technique is useful for correlating local features and makes the model more generalisation; it allows the model to establish a direct correlation by corrupted spatial structure. Hence the finer-grained feature is more emphasised, and the model is more robust. Jigsaw augmentation can be categorised into two groups: full jigsaw (all patches are uniformly shuffled) and partial jigsaw [15]. Ordinary classification (pairs of category labels and images) cannot implement the partial jigsaw directly without localise of ground truth (GT). This ground truth limits the shuffleable area; the area inside the ground truth is unable to shuffle. According to Tao Yu [15], an effect of using the partial jigsaw is improve in F1-score of GIM by 0.9%.

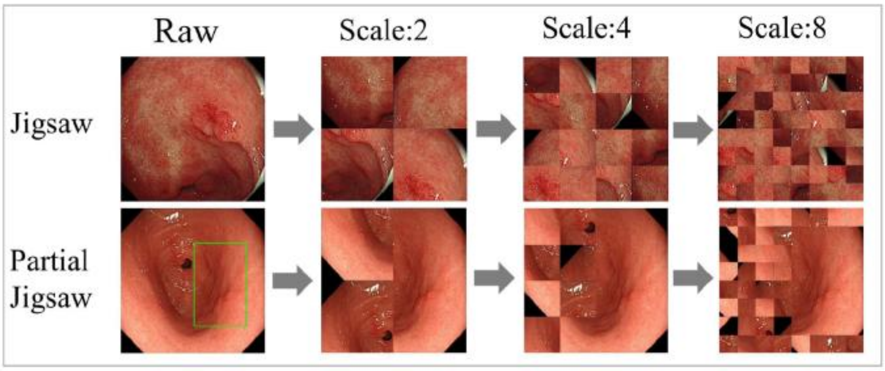


***Figure* 17.** *data augmentation methods: Full Jigsaw and Partial Jigsaw. Partial jigsaw keeps the contents of the GT annotation [15].*

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

3.2.3. Edge-guided path, 2021

The concept of an edge-guided path is to enhance spatial information by enlarging a spatial path of Fast-SCNN with two convolution blocks and Sobel's filter, which provides a guided image [26].

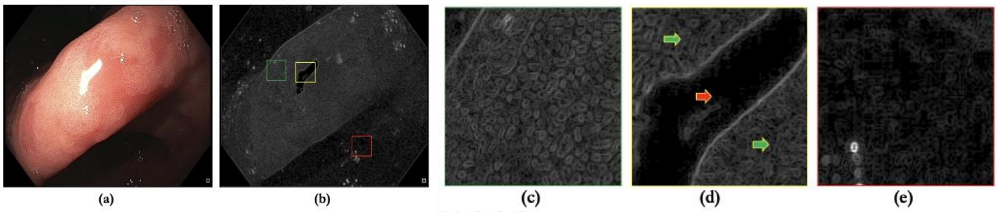


Figure 18. the result of Sobel's edge filter (a) raw image (b) image with Sobel's filter, the cropped images representing: (c) GIM (d) both GIM (the green arrow) and non-GIM (the red arrow) (e) non-GIM [26].

Firstly, for the edge-guided path, RGB images are applied with Sobel's edge filter to transfer normal images to the edge images. Secondly, apply the first convolution block and then add the RGB information from the normal path. Finally, apply the second convolution block and then add the same RGB information; hence, the final guided feature is ready for the regular spatial path of Fast-SCNN (see Figure 18 and 19). A result of deploying an edge-guided path is improved F1-score and IoU by 2.68% and 0.02%.

CHULALONGKORN UNIVERSITY

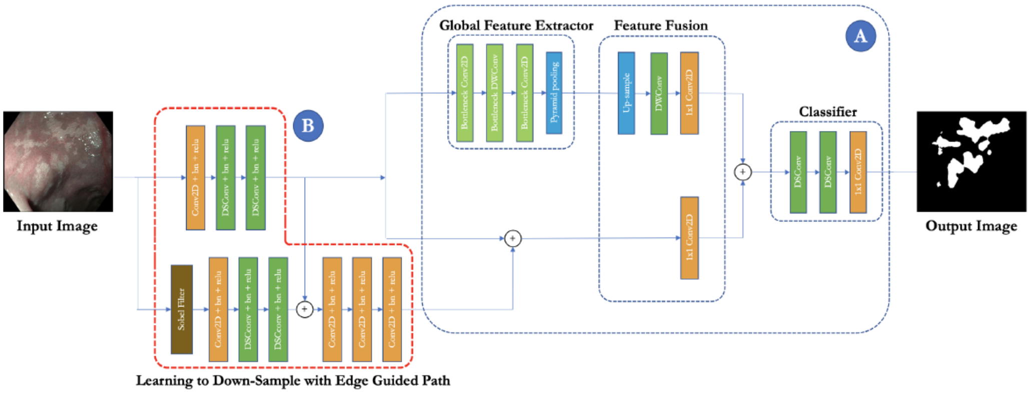


Figure 19. Fast-SCNN modification with edge-guided; (a) core Fast-SCNN model (b) edge-guided block [26].

3.3. Conclusion

All of the relevant neural networks were utilized either as baseline models or backbone models. Since the inference speed and segmentation performance of BiSeNet (the first version), STDC2-Seg50, and BlazeNeo have been notified as effective, they were chosen as backbone models for further modification and employment with other proposed techniques. BiSeNetV2, DDRNet-23-slim, STDC2-Seg50, BlazeNeo, and HRNetV2+OCR [42] were employed as baseline models to benchmark the proposed models and techniques. Furthermore, the jigsaw augmentation for semantic segmentation, one of our proposed strategies, was inspired by Tao Yu [15], the jigsaw augmentation for image classification.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

CHAPTER IV

CONCEPT AND RESEARCH METHODOLOGY

In this chapter, end to end process is explained to be able to perform in real-time with high performance. The research process is as follows: (1) data preparation, (2) image preprocessing, (3) model improvement, (4) model evaluation, and (5) model deployment.

4.1. Data preparation

In this study, endoscopic data were collected from the Center of Excellence for Innovation and Endoscopy in Gastrointestinal Oncology, Chulalongkorn University, Thailand. The dataset included 940 endoscopic images that were annotated as GIM (shown in Figure 20), as well as 1,239 images of non-GIM. Our initial trial showed that there were more instances of the negative class than the positive class, indicating the need to address this imbalance before proceeding. Therefore, we utilized Location-wise hard negative sampling (LW-HNS) to acquire the experimental dataset reported in this study. Additionally, the non-GIM images were divided into three groups: 579 normal EGD images, 515 abnormal EGD images (which were excluded from the GIM category), and 145 noise images (as shown in Figure 21 and Table 1). The LW-HNS method was used to incorporate the abnormal EGD and noise images into the dataset. Expert endoscopists annotated the GIM images using the LabelME tool to identify the GIM legions for our dataset. The annotations were saved in PNG (Portable Network Graphics) format.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

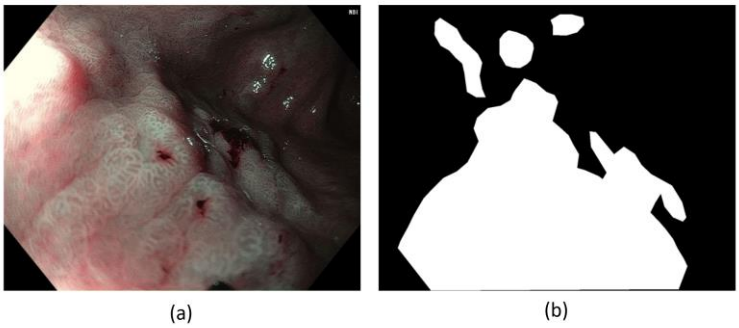


Figure 20. *The sample of (a) GIM image and (b) ground truth.*

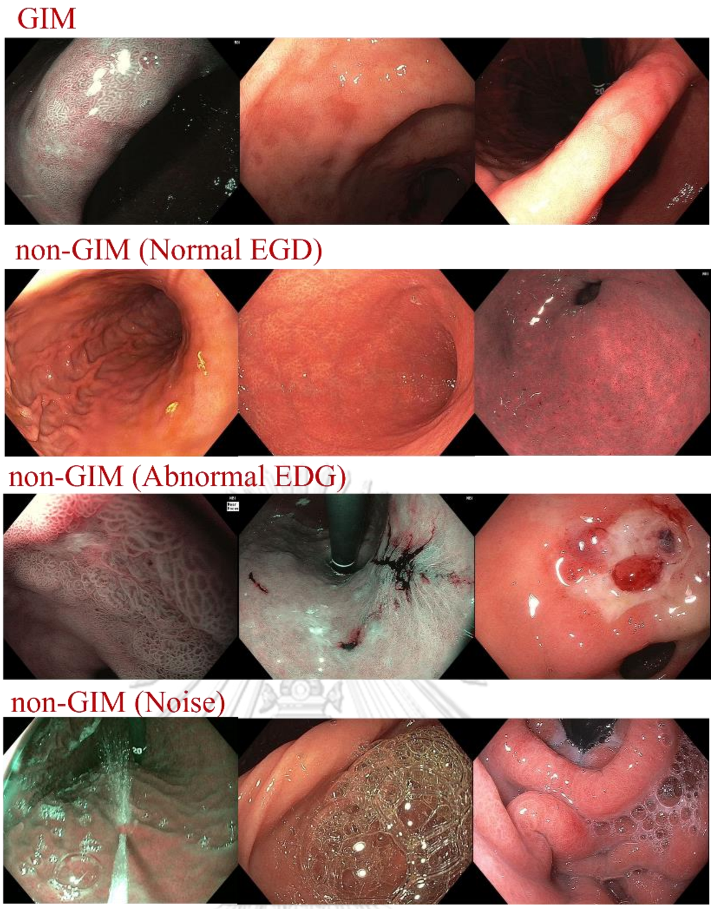


Figure 21. The experimental dataset is comprised of both GIM (positive) and non-GIM (negative) images, with the non-GIM images further classified into three subclasses: normal EGD, abnormal EGD (non-GIM), and noise [28].

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Table 1. The experimental dataset consists of GIM (positive) and non-GIM (negative) images, with the non-GIM images categorized into three subclasses: normal EGD, abnormal EGD (non-GIM), and noise [28].

class	sub-class	images
GIM	-	940
non-GIM	Normal EGD	579
	Abnormal EGD	515
	Noise	145
total	-	2,179

In addition, our dataset is divided into two lighting modes: WLI and NBI, which have varying impacts on the learning process. NBI images contain more distinct features compared to WLE images, resulting in an imbalanced prediction by the model. The number of endoscopic images categorized by the lighting modes is presented in Table 2.

Table 2. *The number of endoscopic images as classified by the lighting modes: NBI and WLE [28].*

	GIM	non-GIM
NBI	565	246
WLE	375	993
total	940	1,239

Table 3 displays the statistics of the data for five different regions of the gastric anatomy: (1) Cardia and fundus in inversion, (2) Corpus in forward view including lesser curve, (3) Corpus in retroflexion including greater curve, (4) Angulus in partial retroflexion, and (5) Gastric antrum.

Table 3. *The number of endoscopic images as classified by gastric anatomical positions [28].*

Gastric Anatomical positions	GIM	non-GIM
Angulus in partial retroflexion	195	128
Cardia and fundus in inversion	4	168
Corpus in forward view including lesser curve	456	494
Corpus in retroflexion including greater curve	87	107
Gastric antrum	198	342
total	940	1,239

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

The GIM images were divided into three sets: a training set consisting of 661 images, a validation set with 93 images, and a testing set containing 186 images. As for the negative samples selected using LW-HNS, 734 images were included in the training set, 111 images in the validation set, and 394 images in the testing set.

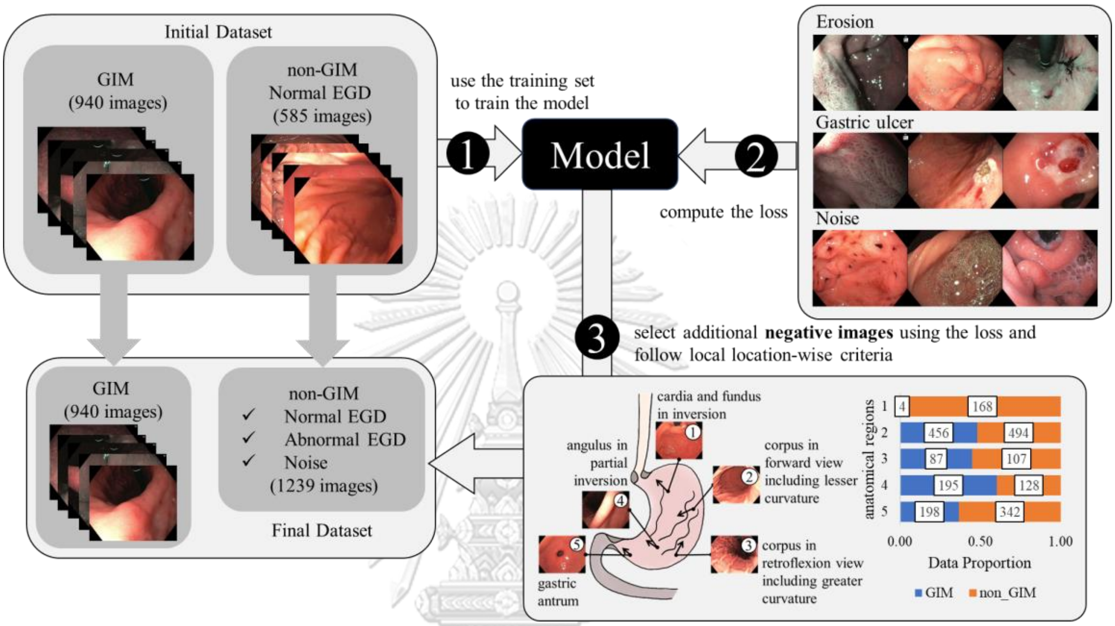


Figure 22. The process of Location-wise hard negative sampling (LW-HNS) comprises three primary steps. Firstly, an initial model is trained using the initial dataset. Secondly, it calculates the loss for negative cases that are not utilized (abnormal EGD and noise images). Lastly, the negative samples are chosen based on their losses in descending order for each category of gastric anatomical positions [28].

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

4.1.1. Location-wise hard negative sampling

The problem of location-wise imbalance arises because certain gastric anatomical positions have a biased distribution of GIM and non-GIM images, which cause predicting bias in the model by the gastric location. To address this issue, a method called location-wise hard negative sampling (LW-HNS) was developed. In most cases, the data provided have a higher number of non-GIM images than GIM images, resulting in an unbalanced proportion of positive and negative samples for each location. This means that negative examples (i.e., non-GIM images) need to be undersampled to achieve a balanced proportion of positive and negative classes.

In unsupervised contrastive learning [43], hard negative sampling (HNS) can help guide the learning process by focusing on negative samples that are in close proximity to the anchor, allowing the model to correct errors more quickly. HNS works by selecting the most difficult negative samples based on their predictions to improve the model's accuracy. In the case of semantic segmentation, after training the initial model with the dataset obtained from [27], the model is used to calculate the loss of the remaining negative images, which include noise and abnormal EGD (such as erosion and ulcer). The remaining negative samples are then sorted based on their loss, and only the top loss negative samples are selected to train the model.

However, simple HNS is inadequate in addressing the issue of location-wise imbalance. Therefore, LW-HNS was developed to balance the data by taking into account the gastric positions. It is necessary to consider the different gastric anatomical positions, as shown in Figure 22. Gastrointestinal endoscopy images can be divided into five categories based on their anatomical location [44]: (1) cardia and fundus in inversion, (2) corpus in forward view including lesser curvature, (3) corpus in retroflexion view including greater curvature, (4) angulus in partial inversion, and (5) gastric antrum. It is important to achieve a balance between positive and negative data for each category, which is added to HNS and referred to as LW-HNS.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

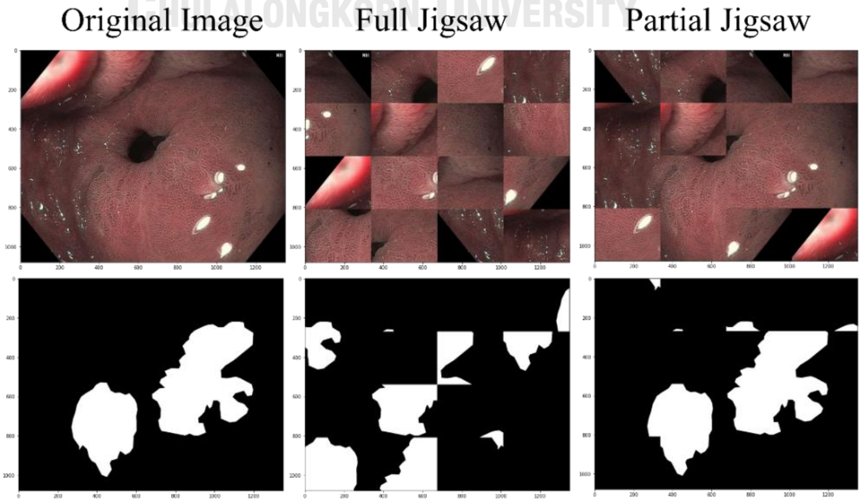


***Figure* 23.** *The full jigsaw and the partial jigsaw augmentation as applied to GIM images [28].*

4.2. Image preprocessing

In the image preprocessing step, the data obtained from section 4.1 is prepared in a suitable format for training the model. Additionally, data augmentation and label smoothing techniques are applied to improve the performance of the model.

4.2.1. Jigsaw augmentation

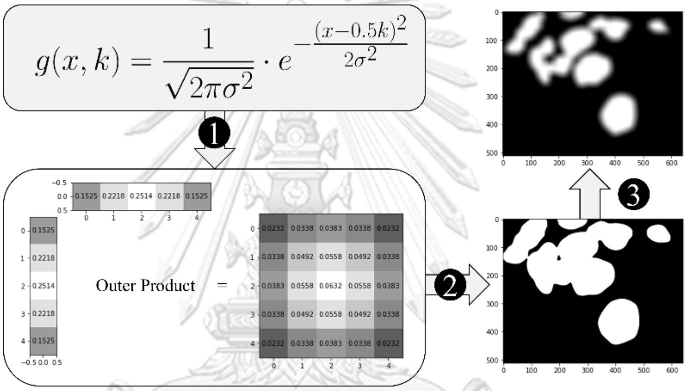
Regular data augmentation and a specific type of augmentation called jigsaw augmentation were applied to the images during preprocessing to prepare them for model training. Jigsaw augmentation was chosen because it can disrupt the spatial structure of the gastric images and help the model establish a strong connection with GIM. The jigsaw technique was originally developed for puzzle problems, but was later found to be useful for DL [45]. The full jigsaw method divides an image into identically sized rectangular pieces, shuffles each piece, and then reassembles them into the original image, as shown in Figure 23. This process enables the model to establish a direct association during training.

Subsequently, a modification of the jigsaw method called "partial jigsaw" was developed to disrupt the spatial structure of gastric images, promote generalization, and disentangle background features from lesion features [15]. Unlike the traditional jigsaw, the partial jigsaw preserves the GIM grids and shuffles only the non-GIM rectangles. This technique is specifically designed for semantic segmentation tasks and has a hyperparameter to determine the GIM grids with a percentage of GIM greater than a predetermined constant. Any other element is considered a non-GIM mesh.

4.2.2. Label smoothing

To improve the training process, the label smoothing technique called gaussian edge softening is used as shown in Figure 24. This method helps to reduce the impact of uncertain edges in the ground truth annotations and allows the model to focus on learning the content within the annotation. By softening the edge, the objective function of the training is relaxed around the edge while emphasizing the inner part of the GIM ground truth.

The gaussian kernel is used to soften the edge, and there are various ways to construct it. One approach involves creating a row and column using finite-state machines (FSMs) and then deriving a convolution kernel from both directions (row and column) [46]. Essentially, the gaussian kernel is produced by taking the outer product of the row and column of gaussian vectors. The values of the gaussian vector can be calculated using the gaussian function, with the center of the gaussian shifted to the middle of the vector.


$g(x,k) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-0.5k)^2}{2\sigma^2}}$
❶
Outer Product
=
❷
❸

	0	1	2	3	4
0	0.1525	0.2218	0.2514	0.2218	0.1525

	0
0	0.1525
1	0.2218
2	0.2514
3	0.2218
4	0.1525

	0	1	2	3	4
0	0.0232	0.0338	0.0383	0.0338	0.0232
1	0.0338	0.0492	0.0558	0.0492	0.0338
2	0.0383	0.0558	0.0632	0.0558	0.0383
3	0.0338	0.0492	0.0558	0.0492	0.0338
4	0.0232	0.0338	0.0383	0.0338	0.0232

Figure 24. *To implement the gaussian edge softening process, first a gaussian vector is created using the gaussian equation. Next, this gaussian vector is used to form the gaussian kernel by taking its outer product with another identical gaussian vector. Finally, this gaussian kernel is applied to the label images to soften their edges and reduce any noisy labels present [28].*

Equation 11 shows the gaussian function, where both the kernel size k of the and the sigma σ corresponding to x. Figure 24 illustrates the process of gaussian edge softening, also known as label smoothing. This process involves applying the gaussian function to both the row and column vectors, followed by defining the kernel as the outer product of these gaussian vectors. Lastly, the obtained kernel is applied to the ground truth image or label:

$$\mathrm{g}(x,k) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-0.5k)^2}{2\sigma^2}} \tag{11}$$

4.2.3. Label smoothing with edge cut

In Figure 25, label smoothing with edge cut is an updated version of the prior version. After softening ground truth with the gaussian kernel, the boundary of the non-zero label is exceeded see Figure 26 (a), causing over-labeling on the negative region. Thus, the previous edge softening label is multiplied by the original ground truth to preserve negative space see Figures 25, the fourth step, and Figures 26 (b).

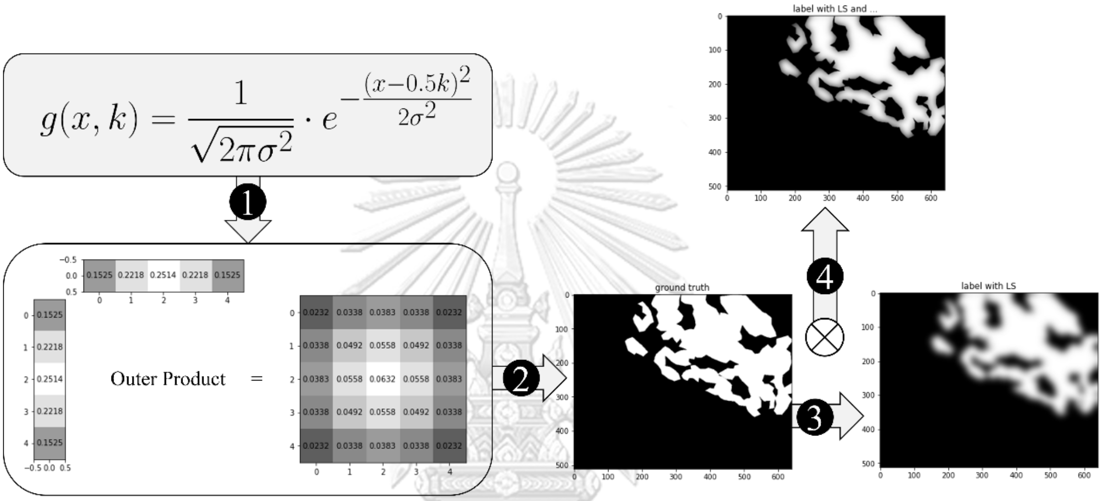
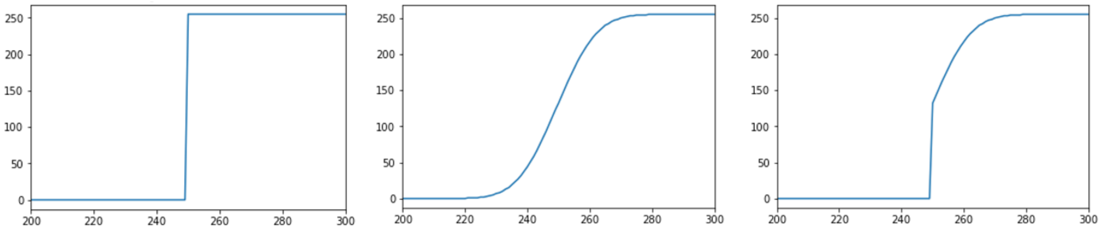


Figure 25. The first, second, and third steps of label smoothing with edge cut remain the same as the previous version. For the fourth, the label with edges softening is multiplied by the original ground truth to limit the non-zero label to being only inside the positive area.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY



(a) Ground truth (b) Label smoothing (c) Label smoothing with edge cut

Figure 26. The cross-section reveals the consequence of involving different label smoothing and the ground truth: (a) ground truth shows zero and positive regions with an edge at 250, (b) regular label smoothing displays over-labeling on the negative side, (c) label smoothing with edge cut preserves the negative site.

4.3. Model improvement

This study attempts to establish a more accurate and practical real-time GIM segmentation for model development. As the best model, the high-speed backbone was selected. In addition, auxiliary head and loss were included to categorise the different colour modes, namely NBI and WLE. In the last step of post-processing, decision thresholds were adjusted independently for NBI and WLE modes, as determined by the auxiliary head.

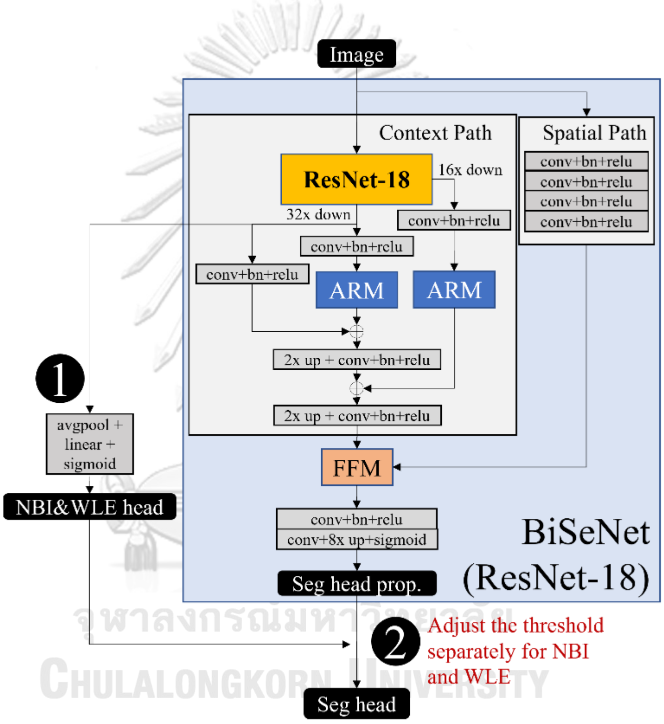
4.3.1. Selective model for GIM image segmentation

Models for GIM semantic segmentation must be able to distinguish the GIM area from healthy mucosa in endoscopic images. In addition, a real-world situation requires an extremely fast inference speed. Thus, in order to accomplish our objective, the optimal model is essential. BiSeNet (with ResNet-18 backbone model), BiSeNetV2, STDC2-seg50, BlazeNeo, DDRNet-23-slim, and HRNetV2 + OCR were candidates for our mission. Although HRNetV2 + OCR is not a real-time image segmentation model, it was used to demonstrate the highest segmentation performance no matter the inference speed.

The selected model must satisfy specific criteria, including average segmentation performance within an image, detection performance for all images in a given dataset, and inference latency. the mean of the intersection over the union region (IoU) between the model prediction and ground truth for GIM images is used to evaluate segmentation performance. Another indicator is "Errors," which can imply unfavourable GIM predictions on non-GIM images. Those IoU and "Error" are employed to determine the GIM detection efficiency. Consequently, the model with the highest F1 score, sensitivity, and negative predictive value (NPV) is selected for the subsequent stage.

4.3.2. Auxiliary head and loss

In order to eradicate imaging mode bias, auxiliary head and loss are added to the tailored model. There are two imaging divisions applicable to EGD: white light endoscopy (WLE) and narrow band imaging (NBI). White light endoscopy (WLE) is an endoscopic image with regular white light. However, WLE has trouble locating GIM due to the unclear texture of GIM. Endoscopists utilise the enhanced image known as NBI for better clarity of GIM textures, thereby increasing the GIM detection rate. This phenomenon occurs when attempting to train a model with a dataset containing a variety of colour modes; consequently, the model is biased by the colour modes. EGD has a tendency to classify NBI images as GIM and WLE images as non-GIM, resulting in diminished prediction power.



จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Figure 27. The enhanced model incorporated an auxiliary head for classifying imaging modes and auxiliary training. In (1), there is the auxiliary head. In (2), decision thresholds can be altered independently for NBI and WLE modes [28].

Figure 27 displays the modifications of the BiSeNet with auxiliary head and loss. The auxiliary path is integrated into the model in order to predict the imaging classes of NBI and WLE. Knowing the difference between NBI and WLE, our model with the auxiliary head is guided to learn the different behaviours for each imaging mode. Furthermore, the auxiliary loss is obtained from the colour-mode classification head. Equation 12 explains how to add an auxiliary loss during the training phase to enable the model to simultaneously segment GIM and differentiate between regular and enhanced images: $loss_{main}$ refers to the cross-entropy loss of GIM segmentation, and $loss_{auxiliary}$ is computed based on the cross-entropy of NBI and WLE classification; w represents the weight of the auxiliary loss. In this experiment, w is equal to 0.1. The auxiliary head is extended from the ResNet-18 backbone model, providing an additional path to the original BiSeNet to enhance the performance of the model:

$$\text{loss}_{total} = \text{loss}_{main} + \text{w} \cdot loss_{auxiliary} \tag{12}$$

4.3.3. Separated threshold adjustment

For every pixel image, the segmentation output of our model provides the GIM confidence as a probability mapping. If confidence exceeds the decision threshold for each pixel, GIM can be assigned; otherwise, the pixel is non-GIM. Generally, a universal threshold was applied to all images, irrespective of imaging mode (NBI or WLE). Thus, the implementation of different thresholds can further enhance the predictive potential. The auxiliary head can provide an alternate classification of NBI and WLE, according to our modified model. This capability allows for applying separate thresholds in each imaging mode.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

4.3.4. Assembling multiple models for NBI and WLE modes

Besides, utilizing an auxiliary head on the single segmentation model to predict imaging modes (NBI and WLE) may corrupt the performances of GIM segmentation, given endoscopic light modes. Since the separated threshold adjustment is vital and model confusion must be avoided, an alternative approach is using another lightweight model, MobileNetV3 [47], instead of the auxiliary head to classify imaging modes. Thus, with the proposed strategy, the segmentation model may deliver better results. Moreover, a fine-tuned model for each color mode is possible. However, the concern drawbacks are error propagation and overall inference time. The models supporting NBI and WLE modes are proposed in 3 approaches (see Figure 28).

Figure 28 illustrates the 3 proposed models that support NBI and WLE modes. Firstly, the single segmentation model with an auxiliary head, which has already been mentioned in section 4.3.2, was designed mainly for BiSeNet model. Secondly, the single segmentation model with the lightweight classification model was presented specifically for STDC2-Seg50 and Blazeneo models. The intuitive idea is that the backbone models (STDC2-Seg50 and Blazeneo) are fine-tuned only for GIM segmentation, leaving NBI and WLE classification tasks for the lightweight model: MobileNetV3. Thirdly, despite fine-tuning the model on both NBI and WLE datasets, the separated fine-tuned model with a lightweight classifier was proposed hopefully to leverage both segmentation and classification for GIM.

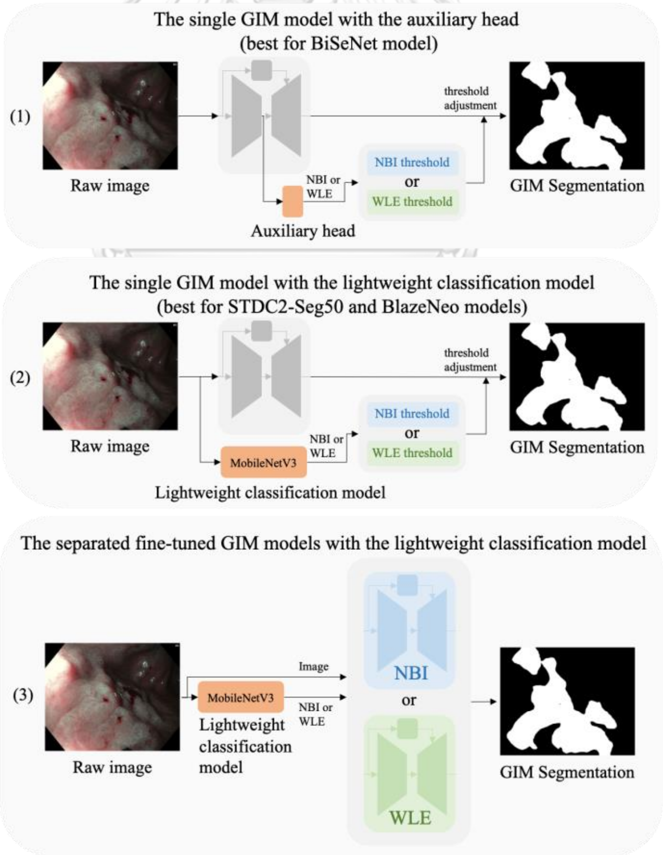

The single GIM model with the auxiliary head
(best for BiSeNet model)
(1)
Raw image
threshold adjustment
NBI threshold
or
WLE threshold
NBI or WLE
Auxiliary head
GIM Segmentation
The single GIM model with the lightweight classification model
(best for STDC2-Seg50 and BlazeNeo models)
(2)
Raw image
threshold adjustment
NBI threshold
or
WLE threshold
NBI or WLE
MobileNetV3
Lightweight classification model
GIM Segmentation
The separated fine-tuned GIM models with the lightweight classification model
(3)
Raw image
MobileNetV3
Lightweight classification model
Image
NBI or WLE
NBI
or
WLE
GIM Segmentation

Figure 28. *The models that supported NBI and WLE modes are proposed in 3 approaches: (1) The single model with the auxiliary head, (2) The single model with*

the lightweight classification model, and (3) the separated fine-tuned models for NBI and WLE modes with the lightweight classification model.

4.4. Model evaluation

The semantic segmentation and classification matrix, as mentioned in 2.4, is utilized for this research. Furthermore, the elemental measurement of speed is also used to measure the inference speed in frames per second. The inference speed is calculated by transforming the original uint8 image to the final uint8 segmentation map (prediction map). Additionally, the proposed techniques, i.e., location-wise hard negative sampling, jigsaw augmentation, label smoothing, auxiliary head, and loss, are applied to the baseline model and some potential networks (BlazeNeo, STDC2-Seg) to confirm the boosting of the model performance. Finally, the error analysis of the best model is included.

4.5. Model deployment

The finalized model will be deployed in open neural network exchange (ONNX) format for practical use in the Center of Excellence in Gastrointestinal Oncology.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

CHAPTER V
EXPERIMENTS AND RESULTS

The experiments and results about Real-time semantic segmentation of GIM are conducted due to the details explained in section 4; The brief navigation is summarized in Table 4.

Table 4. *The brief navigation of the experiments and their winning models or techniques as the sequence.*

Title	The winner	The backbone
5.1. Ablation study for single fine-tuned models using NBI and WLE data.		
Selecting model for GIM image segmentation.	BiSeNet	BiSeNet
Applying location-wise hard negative sampling.	Applied	
Applying jigsaw augmentation.	Applied	
Applying label smoothing.	Applied	
Applying auxiliary head and loss.	Applied	
5.2. Experiment on separated fine-tuned models for NBI and WLE modes.		
Fine-tuned classifier for NBI and WLE class.	MobileNetV3	-
Selecting separated and single models for BiSeNet with MobileNetV3 as a classifier.	single models	BiSeNet
Selecting separated and single models for STDC2-Seg50 with MobileNetV3 as a classifier.	single models	STDC2-Seg50
5.3. Quantitative and qualitative comparisons using the entire dataset.		
Comparison of all proposed strategies for BiSeNet.		BiSeNet
Comparison of all proposed strategies for STDC2-Seg50.		STDC2-Seg50
Comparison of all proposed strategies for BlazeNeo.		BlazeNeo
5.4. Quantitative comparison on abnormal EGD.		
5.5. Error analysis of the winner model.		
5.6. Model deployment.		

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

In Table 4, an ablation study is constructed to find the best combination techniques, followed by an experiment to determine the best between separated or single fine-tuned models. Thirdly, comparisons on the entire dataset are constructed for all backbones to compare the progressive performance of our techniques. Then, a comparison on abnormal EGD is included to examine the model's capability on abnormality. Finally, error analysis and model deployment are displayed.

Table 5. *Performance comparison of different model architectures on the testing set. The LW-HNS technique was applied. Boldface refers to the winner [28].*

Model	Acc	Sen	Spec	F1	PPV	NPV	IoU	Error	FPS
BiSeNet (ResNet-18)	93.37	84.15	**98.06**	89.53	**95.65**	92.41	48.91	**0.2184**	177
BiSeNetV2	84.71	73.22	90.56	76.35	79.76	86.93	41.55	1.3765	**181**
DDRNet-23-slim	89.50	80.33	94.17	83.76	87.50	90.40	40.76	0.4482	98
BlazeNeo	**94.82**	**90.32**	96.67	**91.06**	91.80	**96.04**	**54.65**	0.2789	93
HRNetV2 + OCR	91.18	82.80	95.53	86.52	90.59	91.44	48.29	0.4142	40
STDC2-Seg50	94.19	86.56	97.34	89.69	93.06	94.61	54.35	0.2722	154

5.1. Ablation study for single fine-tuned models on overall data

5.1.1. Selective model for GIM image segmentation

To determine the optimal architecture for our model, six candidates were chosen based on performance and inference speed. Table 5 displays the score for each architecture within a testing set. For real-world deployment, the inference speed must exceed 100. According to this criterion, BiSeNet was chosen as our primary architecture, given that it obtained 89.53% on F1-score and 177 FPS in real-time. Consequently, only BiSeNet was utilised in subsequent experimental sections. Even though BiSeNetV2 achieved the highest speed, other metrics underperformed, especially the F1 score decreased by more than 11%. Even at a higher resolution of 1024x1280, DDRNet-23-slim was inadequate; its inference speed was only 98 FPS. BlazeNeo, on the other hand, outperformed the rest of the models in most categories, except for FPS at 83.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

5.1.2. location-wise hard negative sampling

This investigation displayed the impact of using standard HNS and location-wise HNS (LW-HNS). Since HNS was a technique used to enhance the quality of the training data, it did not slow down the inference speeds of the model. In Table 6, a comparison of HNS strategies for three scenarios is presented. The winning technique was LW-HNS, which had the highest F1 (89.53%). As a result, LW-HNS is observed to improve IoU and F1 scores from HNS by over 4% and 5%, respectively.

Table 6. *Performance comparison of different hard negative sampling strategies on the testing set. Boldface refers to the winner [28].*

Method	Acc	Sen	Spec	F1	PPV	NPV	IoU	Error
No HNS	87.10	**87.57**	87.57	81.79	76.73	**93.39**	**52.94**	2.8365
HNS	92.08	81.42	97.50	87.39	94.30	91.17	44.48	**0.1867**
LW-HNS	**93.37**	84.15	**98.06**	**89.53**	95.65	92.41	48.91	0.2184

5.1.3. Jigsaw augmentation

In Figure 29, the model can distinguish reasonably well between GIM and non-GIM areas, but some structural bias still remains. Bias occurs when the model tries to predict GIM solely based on the gastric morphological features (structures) and ignores the texture features. To address this issue, both full and partial jigsaw augmentations are implemented and compared..

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

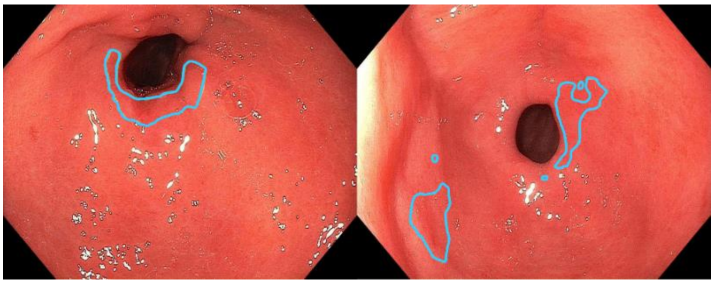


Figure 29. *The instances show non-GIM images that are incorrectly identified as false positives due to structural bias. GIM is likely to appear in some particular areas (structures), such as the hole of gastric autumn, so the prediction is biased by the hole or other gastric structures, regardless of the GIM texture [28].*

This study examined the effects of employing different jigsaw augmentations, including both full and partial jigsaw augmentations. The augmentations were only implemented during training, so they had no effect on inference speed. In Table 7, a comparison of various jigsaw augmentation strategies is presented. The full jigsaw obtained the utmost sensitivity (85.7%) and IoU (49.59%). As a result, the full jigsaw was selected as one of our data pre-processing techniques according to the objective of increasing sensitivity and IoU while maintaining specificity and NPV at or above 90%.

Table 7. *Performance comparison of different jigsaw augmentation strategies on the testing set. Boldface refers to the winner [28].*

Method	Acc	Sen	Spec	F1	PPV	NPV	IoU	Error
without jigsaw	**93.37**	84.15	**98.06**	**89.53**	**95.65**	92.41	48.91	**0.2184**
full jigsaw	93.20	**85.80**	96.93	89.41	93.25	**93.28**	**49.59**	0.3961
partial jigsaw	92.82	84.70	96.94	88.83	93.37	92.57	48.53	0.3314

5.1.4. Label smoothing

In Table 8, label smoothing, also known as gaussian edge softening, can enhance image segmentation and detection efficacy by reducing label noise. The F1 scores of both methods (with and without label smoothing) are greater than 89%. Plus, this method can increase sensitivity and IoU by nearly 1.2% and 4%, respectively.

Table 8. *An effect of our label smoothing (gaussian edge smoothing) on the testing set. Boldface refers to the winner [28].*

Method	Acc	Sen	Spec	F1	PPV	NPV	IoU	Error
without noisy label	**93.20**	85.80	**96.93**	**89.41**	**93.25**	93.28	49.59	0.3961
with gaussian edge softening	92.90	**87.01**	95.81	89.02	91.12	**93.72**	**53.53**	**0.3108**

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

5.1.4. Auxiliary head and loss

Multiple image modes, namely white light endoscopy (WLE) and narrow-band images (NBI), are supported by auxiliary head and loss. Table 9 demonstrates that using the auxiliary head and loss together with different threshold adjustments produces outstanding outcomes. The model supports GIM segmentation and imaging mode classification, including NBI and WLE. The auxiliary head can also be used for altering the GIM-confidential threshold for different lighting modes with little effect on error propagation. Generally, a universal threshold was applied to all images, regardless of the imaging mode (NBI or WLE). Figure 30 demonstrates that the model tends to set up different confidence intervals for NBI and WLE. The NBI confidence is larger than the WLE confidence. Thus, a different threshold adjustment boosts the performance of the model. In the overall section, all matrices achieved a minimum of 90%. The sensitivity is raised to 91.40%, and the F1 score is enhanced to 91.15%. In the WLE section, sensitivity is increased by roughly 10%, and the F1 score is improved by around 83%. NBI efficiency is boosted in the NBI section by about 17% specificity, affecting the F1 score by 3%.

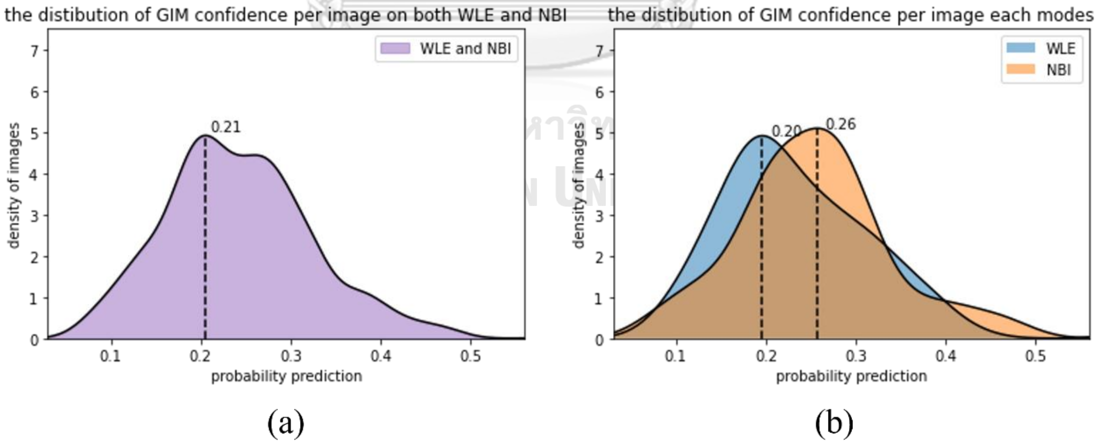


Figure 30. *The distribution of GIM confidence per image (taking an average for all pixels) on the validation set: (a) all images contain imaging modes and (b) separated imaging modes. Since NBI is more confident than WLE, it is preferable to use separate thresholds in this case [28].*

Table 9. *The effect of adding auxiliary head and loss into the network along with threshold adjustment on the testing set. There are three parts in the Table 8: (1) overall images, (2) WLE images, and (3) NBI images. Boldface refers to the winner [28].*

Method	Acc	Sen	Spec	F1	PPV	NPV	IoU	Error	FPS
overall modes									
without auxiliary head	92.90	87.01	95.81	89.02	91.12	93.72	53.53	**0.3108**	**177**
with auxiliary head (universal threshold)	**94.31**	90.86	**95.94**	91.11	**91.35**	95.70	54.09	0.4587	173
with auxiliary head (separate thresholds)	**94.31**	**91.40**	95.69	**91.15**	90.91	**95.93**	**54.66**	0.5256	173
WLE mode									
without auxiliary head	92.96	75.76	**96.24**	77.52	79.37	95.42	48.76	**0.3050**	**177**
with auxiliary head (universal threshold)	94.17	84.85	95.95	82.35	80.00	97.08	48.48	0.4958	173
with auxiliary head (separate thresholds)	**94.42**	**86.36**	95.95	**83.21**	**80.28**	**97.36**	**49.60**	0.5691	173
NBI mode									
without auxiliary head	89.88	**94.17**	79.17	93.00	91.87	84.44	57.57	1.0753	**177**
with auxiliary head (universal threshold)	**94.64**	**94.17**	**95.83**	**96.17**	**98.26**	**86.79**	57.18	**0.1915**	173
with auxiliary head (separate thresholds)	94.05	**94.17**	93.75	95.76	97.41	86.54	**57.45**	0.2117	173

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

5.2. Separated fine-tuned models for NBI and WLE modes with a classifier

Intuitively, fine-tuned models for NBI and WLE modes with the classification model provide more dynamics to the overall performance by trading off inference speed. In Table 11, the performance of BiSeNet model demonstrates that utilizing separated models with NBI&WLE classification model (MobileNetV3s) impairs F1-score, IoU by 0.74%, 0.48% relative to the single model with an auxiliary head. Additionally, the inference speed dramatically declined from 173 FPS to 135 FPS. Thus the performance of the fine-tuned models cannot compare to the solo model with the auxiliary head, even though the performance of the classification model on categories imaging modes is exceptionally high, which achieved an F1-score of 99.88% in 1.73 ms (see Table 10). In NBI modes, the fine-tuned model is superior, achieving 97.02% of the F1-score; meanwhile, the performance in WLE modes is dropped from 83.21% to 78.46% of the F1-score, compared to the solo model with an auxiliary head. In Table 12, the experimental result demonstrates that applying the same techniques to the STDC2-Seg50 effect in a similar trend. The separately fine-tuned models are overwhelmed by the single model with the classifier in both NBI and WLE modes.

Table 10. *Performance of NBI&WLE classification model using MobileNetV3s and speed in milliseconds (ms).*

Model	Acc	Sen	Spec	F1	PPV	NPV	ms
MobileNetV3s	99.83	**100.00**	99.76	99.88	99.41	100.00	1.73

In Figure 31, the WLE fine-turned model tends to shrink the segmentation region in GIM areas; unfortunately, it reduces the similarity between the ground truths and the predictions causing lower performance relative to the universal model with the auxiliary head. Another reason is that the solo model utilizes the entire training dataset, causing more diversity and guidance. The GIM in NBI has more explicit texture than GIM in WLE; using only WLE cause the model to be confused about establishing a direct correlation between the prediction with GIM features: texture and color.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

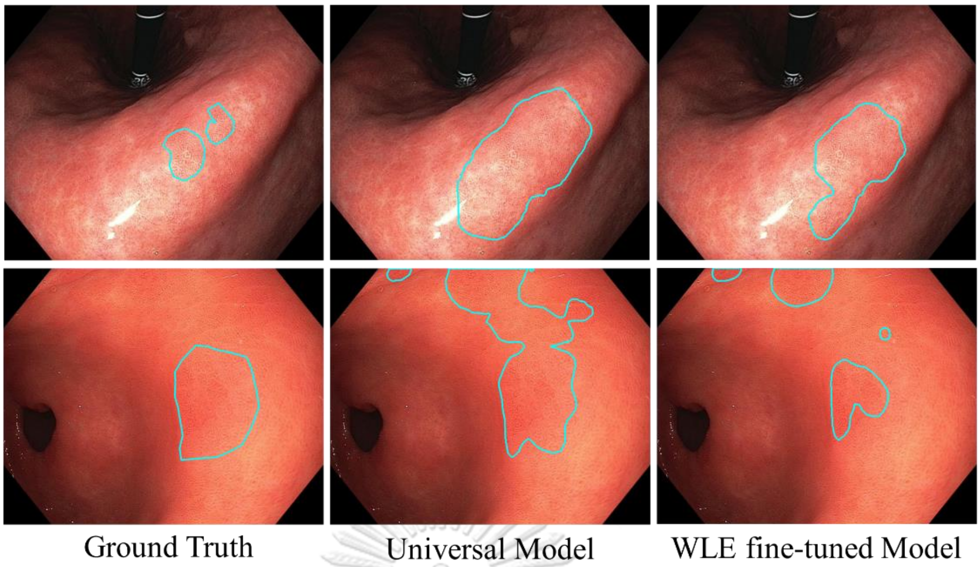


Figure 31. The output of the WLE fine-tuned model is overshot compared to the ground truth, yielding a lower IoU score relative to the universal model with an auxiliary head.

Table 11. Performance comparison between a single BiSeNet model with an auxiliary head (separate thresholds), and separated fine-tuned BiSeNet models for NBI and WLE modes plus the classification model (MobileNetV3s). There are three parts in the Table 10: (1) overall images, (2) WLE images, and (3) NBI images. Boldface refers to the winner.

Method	Acc	Sen	Spec	F1	PPV	NPV	IoU	Error	FPS
overall modes									
single model	**94.31**	**91.40**	95.69	**91.15**	90.91	**95.93**	**54.66**	0.5256	**173**
separated models	93.97	88.71	**96.45**	90.41	**92.18**	94.76	54.18	**0.3322**	135
WLE mode									
single model	**94.42**	**86.36**	95.95	**83.21**	80.28	**97.36**	**49.60**	0.5691	**173**
separated models	93.20	77.27	**96.24**	78.46	**79.69**	95.69	46.75	**0.3348**	135
NBI mode									
single model	94.05	94.17	93.75	95.76	97.41	86.54	57.45	**0.2117**	**173**
separated models	**95.83**	**95.00**	**97.92**	**97.02**	**99.13**	**88.68**	**58.82**	0.5554	135

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Table 12. *Performance comparison between a single STDC2-Seg50 model with NBI&WLE classifier (MobileNetV3s and separate thresholds) and separated fine-tuned STDC2-Seg50 models for NBI&WLE modes plus the classifier (MobileNetV3s). There are three parts in the Table 10: (1) overall images, (2) WLE images, and (3) NBI images. Boldface refers to the winner.*

Method	Acc	Sen	Spec	F1	PPV	NPV	IoU	Error	FPS
overall modes									
single model	**94.98**	90.86	**96.67**	**91.35**	**91.85**	96.25	**57.34**	**0.2512**	121
separated models	94.03	**91.40**	95.12	89.95	88.54	**96.40**	53.42	0.3512	121
WLE mode									
single model	**94.17**	81.82	**96.53**	**81.82**	**81.82**	**96.53**	**52.43**	**0.1860**	121
separated models	93.69	**83.33**	95.66	80.88	78.57	96.78	46.21	0.2348	121
NBI mode									
single model	96.44	**95.83**	**97.14**	**96.64**	**97.46**	**95.33**	**60.05**	**0.4664**	121
separated models	**94.67**	**95.83**	93.33	95.04	94.26	95.15	57.38	0.7346	121

5.3. Comparison on the entire dataset

5.3.1. Quantitative comparison

Every experiment has been recapped. A quantitative comparison of our model is presented in Table 13. Our model with all strategies combined outperformed BlazeNeo for default training on the GIM dataset in terms of speed and sensitivity, NPV, and IoU. It achieved 91.40% sensitivity, 95.93% NPV, 54.66% IoU, and 173 FPS. In Tables 13 and 16, BiSeNet represents the original model with an enhancement (pre-processing and post-processing) proposed in our previous work [28]. Moreover, since implementing LW-HNS, our model has become more robust regarding abnormal EGD (Table 16), demonstrating a substantial improvement over the original BiSeNet from 69.92% to more than 90%.

In Table 14, the quantitative comparison of STDC2-Seg50 revealed progressive performance via the proposed strategies, including the classifier of MobileNetV3s. Location-wise hard negative sampling combined with the full jigsaw augmentation achieved the best F1-score for STDC2-Seg50 model at 93.48% with a speed of 154

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

FPS. However, when utilizing label smoothing with edge cut, IoU was improved by 1.67% from 55.68% to 57.35%, whilst F1-score was impaired from 93.48% to 91.35%. After applying the NBI&WLE classifier, the performance was not significantly changed from the previous one, but the inference speed declined from 154 FPS to 121 FPS.

BlazeNeo, combined with all proposed techniques, shared a similar story as STDC2-Seg50. In Table 15, BlazeNeo, with all proposed strategies and the NBI&WLE classifier, achieved the best IoU at 58.30% triumph over all our models regardless of inference speed. Regarding the inference speed, BlazeNeo+LW-HNS+FJ+LS reached the best F1-score of 93.44% with a speed of 93 FPS.

In summary, all the proposed strategies have the tendency to increase the performance of GIM semantic segmentation in Real-Time relative to the default models of BiSeNet, STDC2-Seg50, and BlazeNeo. The best outcome varied based on the nature of the default models: two-path-way architecture and single-path-way architecture. Nonetheless, our approach leveraged the maximum capabilities of the baseline models with a slight impact on the inference speeds.

5.3.2. Qualitative comparison

Figure 32 illustrates how each proposed technique, including LW-HNS, the full jigsaw augmentation (FJ), label smoothing (LS), and auxiliary head (AuxHead), contributes to output prediction. There are 12 samples displayed here: the 1st - 4th rows contain NBI GIM images, the 5th - 8th rows contain WLE GIM images, the 9th - 10th rows contain NBI non-GIM images, and the 11th - 12th rows contain WLE non-GIM images. As for the column, 1st and 2nd columns are the original and ground truth images. The remaining columns are our prediction outcomes; therefore, the last column is the victor applying all proposed techniques. Interestingly, the initial BiSeNet performs inadequately on non-GIM images (9th to 12th rows: column (c)).

Figure 33 illustrates the best result of each model via the proposed techniques and its baseline: BiSeNet, STDC2-Seg50, and BlazeNeo. Note that all baseline models were applied LW-HNS, and the 12 instances are the same as in Figure 32. As for the column, the 1st and 2nd are the same, whilst the remaining columns contain prediction outcomes. The last column reveals that BlazeNeo integrated with all strategies outperformed other models regardless of an error in abnormal EGD, 9th row. Curiously, our method improved the performance of all baseline models for all predictions.

5.4. Quantitative comparison on Abnormal EGD

A comparison of performance on 133 abnormal EGD cases showed that the prediction efficiency (accuracy) of BiSeNet and STDC2-Seg50 improved from 69.92% to 92.48% and 44.36% to 96.24%, respectively, see Table 16. Furthermore, the winner of BlazeNeo achieved the best accuracy at 97.74% regardless of inference speed.

5.5. Error Analysis of the winner model with the best techniques

Among the three enhanced models (BiSeNet, STDC2-Seg50, and BlazeNeo), STDC2-Seg50 with location-wise hard negative and full jigsaw augmentation is preferable due to its performance and the inference speed exceeding 100 FPS at 154 FPS. Additionally, the speed was impaired from BiSeNet by 23 FPS from 177 FPS to 154 FPS; meanwhile, BlazeNeo's speed dramatically dropped to 80 FPS. Thus STDC2-Seg50 was chosen. The error analysis of the STDC2-Seg50 is as follows:

5.5.1. Error Analysis of False Negative segmentation

According to the classification performance of STDC2-Seg50 with LW-HNS and full jigsaw augmentation, 14 of 186 GIM images were incorrectly classified as negative images. Six of them were NBI images, and the other was WLE images. Herein, the recall of GIM in NBI and WLE mode was 95.00% and 87.88%, meaning that WLE was more errors than NBI mode. In NBI, most errors were incisura (recall 89.26% GIM); see Figure 34. These occurred since we tried to reduce the massive false positive (FP) error by adding more negative data; the FP is critical. For WLE, the incisura images are the majority of false negative images for the same reason. Moreover, the error emerged on the GIM image that does not have the precise pattern of GIM (too complicated and easy to overlook); see Figure 35.

Table 13. *Quantitative comparison of BiSeNet model showing each stage of improvements: location-wise hard negative sampling (LW-HNS), the full jigsaw augmentation as (FJ), label smoothing (LS), and auxiliary head (AuxHead). Our model is also compared with the baseline. Boldface refers to the winner [28].*

Method	Acc	Sen	Spec	F1	PPV	NPV	IoU	Error	FPS
BiSeNet [28]	87.10	87.57	86.87	81.79	76.73	93.39	51.94	1.8365	177
BiSeNet+LW-HNS	93.37	84.15	**98.06**	89.53	**95.65**	92.41	48.91	**0.2184**	177
BiSeNet+LW-HNS+FJ	93.20	85.80	96.93	89.41	93.25	93.28	49.59	0.3961	177
BiSeNet+LW-HNS+FJ+LS	92.90	87.01	95.81	89.02	91.12	93.72	53.53	0.3108	177
BiSeNet+LW-HNS+FJ+LS+AuxHead	**94.31**	**91.40**	95.69	**91.15**	90.91	**95.93**	**54.66**	0.5256	173
BiSeNetV2	84.71	73.22	90.56	76.35	79.76	86.93	41.55	1.3765	**181**
DDRNet-23-slim*	89.50	80.33	94.17	83.76	87.50	90.40	40.76	0.4482	98
BlazeNeo	94.82	90.32	96.67	91.06	91.80	96.04	54.65	0.2789	93
HRNetV2 + OCR	91.18	82.80	95.53	86.52	90.59	91.44	48.29	0.4142	40
STDC2-Seg	94.19	86.56	97.34	89.69	93.06	94.61	54.35	0.2722	154

LW-HNS has been applied to BiSeNetV2, DDRNet-23-slim, BlazeNeo, HRNetV2+OCR, and STDC2-Seg.

* It was implemented on 1024x1280 image resolution.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Table 14. *Quantitative comparison of STDC2-Seg50 model showing each stage of improvements: location-wise hard negative sampling (LW-HNS), the full jigsaw augmentation as (FJ), label smoothing with edge cut (LS), auxiliary head (AuxHead), and NBI&WLE classifier (MobileNetV3s). Our model is also compared with the baseline. Boldface refers to the winner.*

Method	Acc	Sen	Spec	F1	PPV	NPV	IoU	Error	FPS
STDC2-Seg50	77.24	91.94	70.30	72.15	59.38	94.86	55.26	3.3984	154
STDC2-Seg50+LW-HNS	93.79	86.56	97.21	89.94	93.06	93.87	54.35	0.3030	154
STDC2-Seg50+LW-HNS+FJ	**95.86**	**92.47**	**97.46**	**93.48**	**94.51**	**96.48**	55.68	**0.1853**	154
STDC2-Seg50+LW-HNS+FJ+LS	94.48	90.86	96.19	91.35	91.85	95.71	**57.35**	0.2866	154
STDC2-Seg50+LW-HNS+FJ+LS+AuxHead	94.35	87.63	97.12	90.06	92.61	95.01	52.53	0.3889	153
STDC2-Seg50+LW-HNS+FJ+LS+MobileNetV3s	94.98	90.86	96.67	91.35	91.85	96.25	57.34	0.2512	121
BiSeNet**	94.31	91.40	95.69	91.15	90.91	95.93	54.66	0.5256	173
BiSeNetV2	84.71	73.22	90.56	76.35	79.76	86.93	41.55	1.3765	**181**
DDRNet-23-slim*	89.50	80.33	94.17	83.76	87.50	90.40	40.76	0.4482	98
BlazeNeo	94.82	90.32	96.67	91.06	91.80	96.04	54.65	0.2789	93
HRNetV2 + OCR	91.18	82.80	95.53	86.52	90.59	91.44	48.29	0.4142	40

LW-HNS has been applied to BiSeNetV2, DDRNet-23-slim, BlazeNeo, and HRNetV2+OCR.

* It was implemented on 1024x1280 image resolution.

** the best version of BiSeNet applying LW-HNS, FJ, LS, and AuxHead.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Table ***15.*** *Quantitative comparison of BlazeNeo (DHA) model showing each stage of improvements: location-wise hard negative sampling (LW-HNS), the full jigsaw augmentation as (FJ), label smoothing (LS), and NBI&WLE classification model (MobileNetV3s). Our model is also compared with the baseline. Boldface refers to the winner [28].*

Method	Acc	Sen	Spec	F1	PPV	NPV	IoU	Error	FPS
BlazeNeo+LW-HNS	94.82	90.32	96.67	91.06	91.80	96.04	54.65	0.2789	93
BlazeNeo+LW-HNS+FJ	95.76	**93.55**	96.67	92.80	92.06	97.32	58.02	0.2610	93
BlazeNeo+LW-HNS+FJ+LS	**96.23**	91.94	98.00	**93.44**	95.00	**96.72**	58.26	0.2506	93
BlazeNeo+LW-HNS+FJ+LS+AuxHead	95.29	89.25	97.78	91.71	94.32	95.66	54.16	0.2750	93
BlazeNeo+LW-HNS+FJ+LS+MobileNetV3s	**96.23**	91.40	**98.23**	93.41	**95.50**	96.51	**58.30**	**0.2253**	80
BiSeNet**	94.31	91.40	95.69	91.15	90.91	95.93	54.66	0.5256	173
BiSeNetV2	84.71	73.22	90.56	76.35	79.76	86.93	41.55	1.3765	**181**
DDRNet-23-slim*	89.50	80.33	94.17	83.76	87.50	90.40	40.76	0.4482	98
STDC2-Seg50***	96.08	90.86	98.23	93.11	95.48	96.30	55.57	0.1446	154
HRNetV2 + OCR	91.18	82.80	95.53	86.52	90.59	91.44	48.29	0.4142	40

LW-HNS has been applied to BiSeNetV2, DDRNet-23-slim, BlazeNeo, and HRNetV2+OCR.

* It was implemented on 1024x1280 image resolution.

** the best version of BiSeNet applying LW-HNS, FJ, LS, and AuxHead.

*** the best version of STDC2-Seg50 applying LW-HNS, and FJ.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

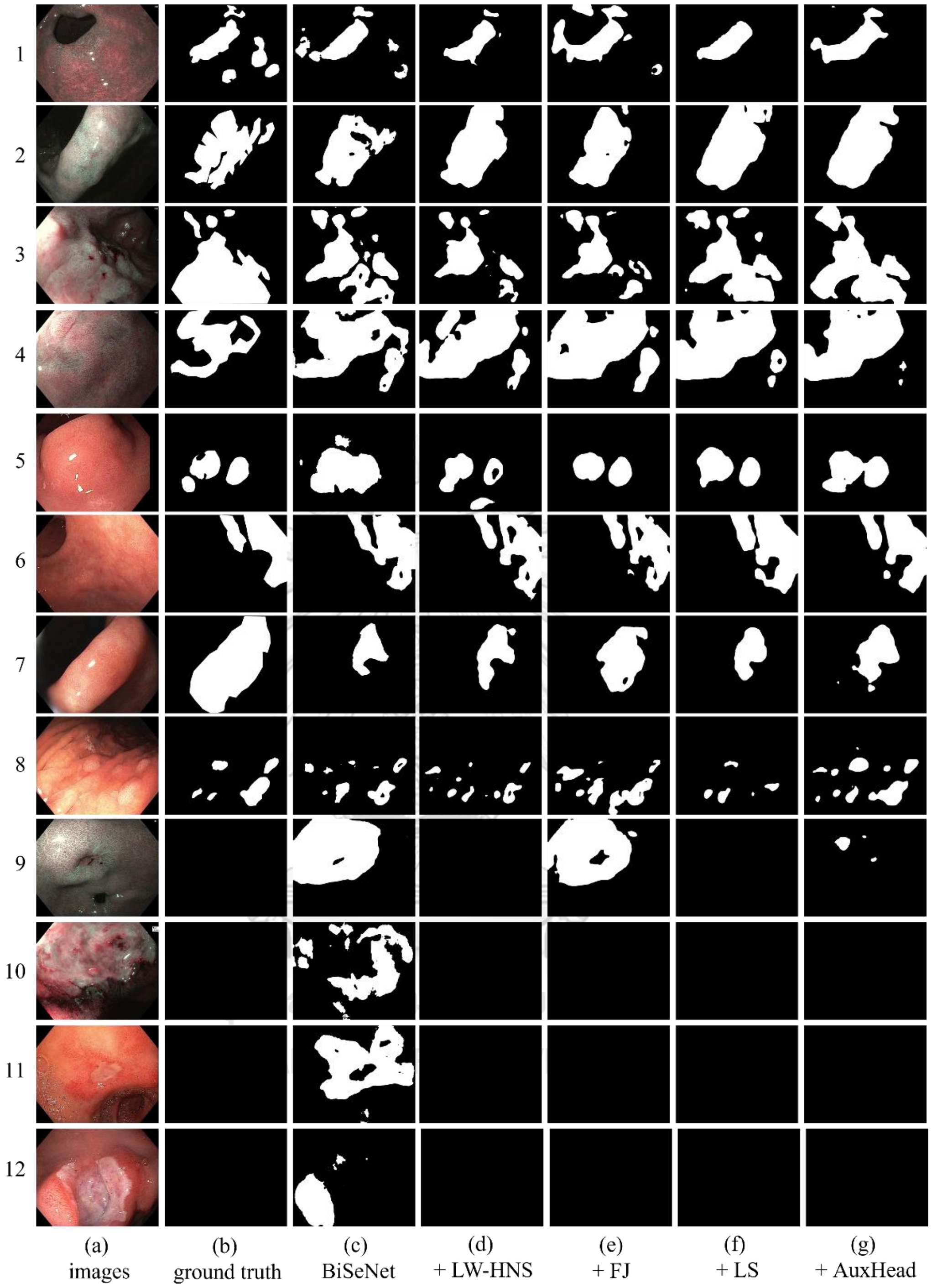


Figure 32. *On the test dataset, a qualitative comparison of the proposed method for BiSeNet (ResNet-18) is illustrated [28]: (a) original image; (b) ground truth; (c) BiSeNet; (d) BiSeNet+LW-HNS; (e) BiSeNet+LW-HNS+FJ; and (g) BiSeNet+LW-HNS+FJ+LS+AuxHead. There are 12 instances (rows) of NBI GIM (1st–4th), WLE GIM (5th–8th), NBI non–GIM with abnormal (9th–10th), and WLE non–GIM with abnormal (11th–12th). (c) cite the preceding work [28].*

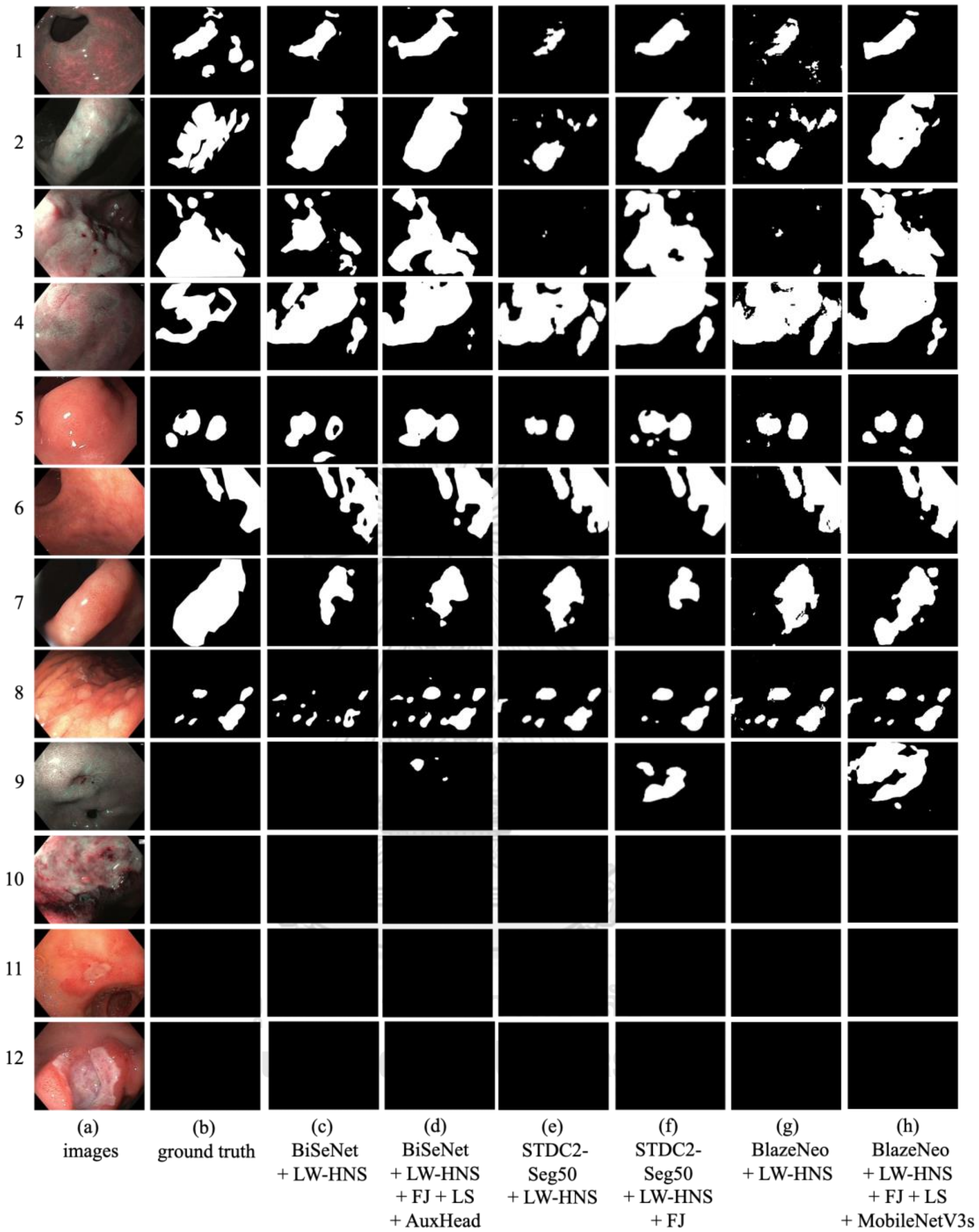

1
2
3
4
5
6
7
8
9
10
11
12
(a) images
(b) ground truth
(c) BiSeNet + LW-HNS
(d) BiSeNet + LW-HNS + FJ + LS + AuxHead
(e) STDC2-Seg50 + LW-HNS
(f) STDC2-Seg50 + LW-HNS + FJ
(g) BlazeNeo + LW-HNS
(h) BlazeNeo + LW-HNS + FJ + LS + MobileNetV3s

***Figure* 33.** *Qualitative comparison of the winner method for each model and the original baseline model on the test dataset: (a) original image, (b) ground truth, (c) BiSeNet+LW-HNS, (d) BiSeNet+LW-HNS+FJ+LS+AuxHead, (e) STDC2-Seg50+LW-HNS, (f) STDC2-Seg50+LW-HNS+FJ, (g) BlazeNeo+LW-HNS, (h) BlazeNeo+LW-HNS+FJ+LS +MobileNetV3s. There are 12 examples (rows) composed of NBI GIM (1st-4th), WLE GIM (5th-8th), NBI non-GIM with abnormal (9th-10th), and WLE non-GIM with abnormal (11th-12th).*

Table 16. *Performance comparison of 133 abnormal EGD cases. Boldface refers to the winner from Table 13, 14, and 15.*

Model	TN	FP	Acc
BiSeNet [28]	93	40	69.92
BiSeNet+LW-HNS	124	9	93.23
BiSeNet+LW-HNS+FJ	126	7	94.74
BiSeNet+LW-HNS+FJ+LS	126	7	94.74
BiSeNet+LW-HNS+FJ+LS+AuxHead	**123**	**10**	**92.48**
STDC2-Seg	59	74	44.36
STDC2-Seg+LW-HNS	128	5	96.24
STDC2-Seg+LW-HNS+FJ	**128**	**5**	**96.24**
STDC2-Seg+LW-HNS+FJ+LS	129	4	96.99
STDC2-Seg+LW-HNS+FJ+LS+AuxHead	124	9	93.23
BlazeNeo+LW-HNS	127	6	95.49
BlazeNeo+LW-HNS+FJ	127	6	95.49
BlazeNeo+LW-HNS+FJ+LS	130	3	97.74
BlazeNeo+LW-HNS+FJ+LS+MobileNetV3s	**130**	**3**	**97.74**
BiSeNetV2	113	20	84.96
DDRNet-23-slim*	120	13	90.23
HRNetV2+OCR	125	8	93.98

LW-HNS has been applied to BiSeNetV2, DDRNet-23-slim, and HRNetV2+OCR.

* It was implemented on 1024x1280 image resolution.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

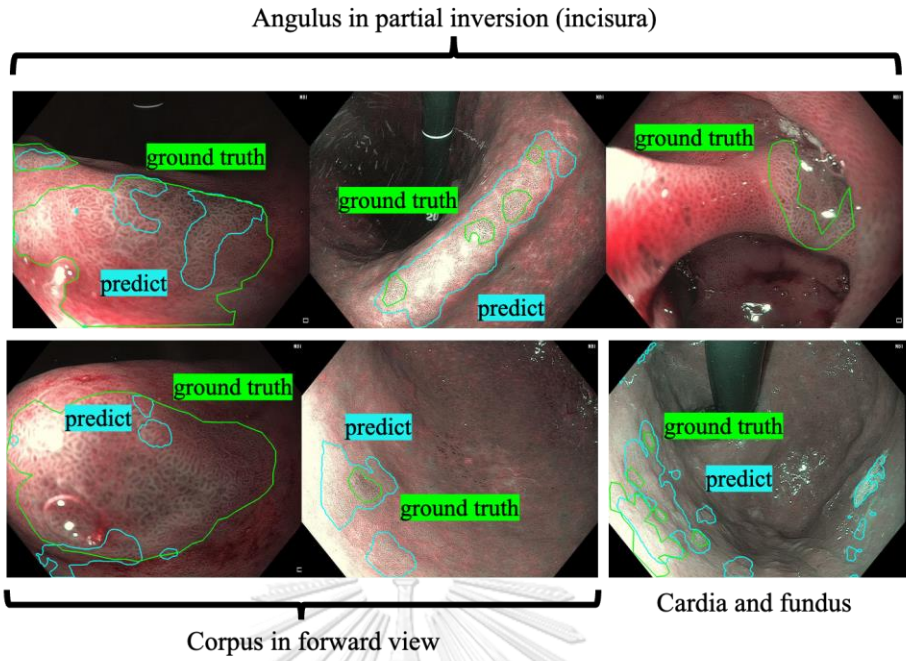


Figure 34. *The false negative images of GIM in NBI mode: 3 images of incisura, 2 images of corpus, and 1 image of cardia and fundus.*

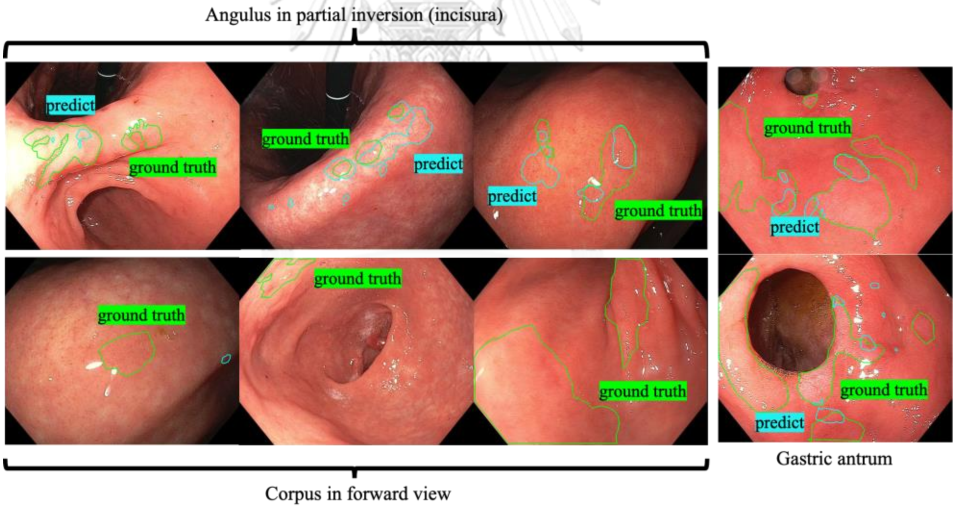


Figure 35. *The false negative images of GIM in WLE mode: 3 images of incisura, 3 images of corpus, and 2 images of antrum. Some images can be overlooked easily.*

5.5.2. Error Analysis of False Positive segmentation.

The False Positive for the STDC2-Seg50 was 10 of 374 non-GIM images, 3 of them were NBI images, and the rest were WLE. In NBI, there were 2 images of abnormal EGD (erosion and ulcer) and one of normal EGD; see Figure 36. For WLE, there were 3 images of abnormal EGD, 3 images of normal EGD, and one of noise; see Figure 37. Here, the number of abnormal EGD indicated that there are some textures of GIM that the model still confuses. Interestingly, the false positive error of normal EGD occurred near the edge of the image; it appeared when the endoscope was near the gastric surface, revealing some patterns of the healthy surface. Perhaps, the model has some camera region bias (top right, top left, etc.).

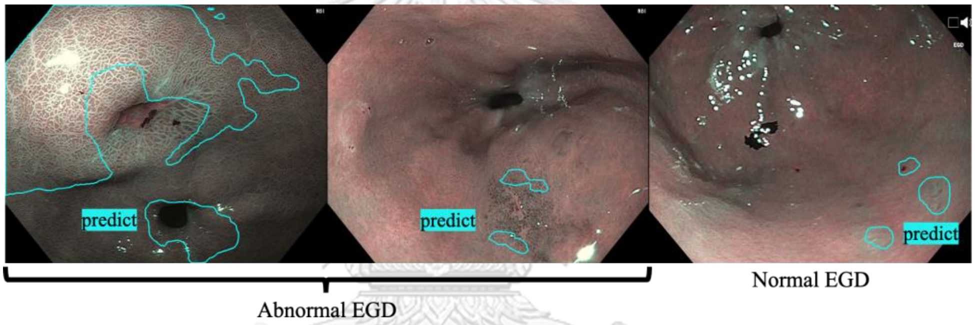


***Figure* 36.** *The false positive images of GIM in NBI mode: 2 images of abnormal EGD and one of normal EGD.*

จุฬาลงกรณ์มหาวิทยาลัย

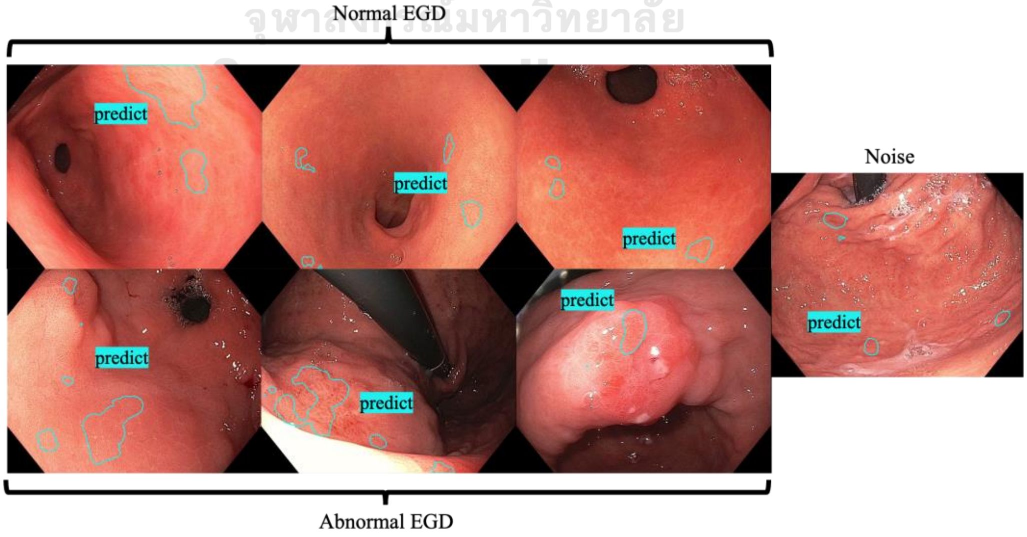


***Figure* 37.** *The false positive images of GIM in WLE mode: 3 images of abnormal EGD, 3 images of normal EGD, and one noise.*

5.6. model deployment

This section presents the performance of the best GIM model of BiSeNet series when deployed since the model can operate without lacking. Here is the first video available on YouTube and the published paper, https://youtu.be/GiG5MCxJh20. The first 8 seconds demonstrated the performance of GIM segmentation in Real-Time on WLE mode, and at the 24 seconds to the end of the clip presented the performance on narrow-band imaging (NBI). The second video, online link – https://youtu.be/B_mtmvZAIbE, demonstrated the AI model on normal EGD and GIM. In the first 30 seconds, the effectiveness of the AI on normal EGD with white light endoscopy is displayed. Between 0:30 to 0:52, the clips showed the AI performing on normal EGD with NBI. Finally, the duration of 0:52 to 1:80 and 1:80 to the end revealed the performance of GIM segmentation on WLE and NBI modes, respectively.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

CHAPTER VI

CONCLUSION

In this thesis, a real-time semantic segmentation network for GIM was designed to conquer all practical obstacles. By adding an auxiliary head and loss, the BiSeNet model was improved. Several techniques were introduced to enhance the model's performance, such as location-wise hard negative sampling, jigsaw augmentation, label smoothing, and threshold adjustment. Our model outperformed all baseline methods (BiSeNet, BiSeNetV2, DDRNet, BlazeNeo, and HRNetV2 + OCR) with F1 and IoU scores of 91.15 and 54.66%, respectively. The inference speed attained 173 FPS. Comparing the performance of 133 abnormal EGD cases revealed an increase in prediction efficiency (accuracy) from 69.92% to 92.40%. In addition, the proposed methodologies have been demonstrated to be effective and more practical on both baseline models: STDC2-Seg50 and BlazeNeo. The performance of STDC2-Seg50 was improved and reached 93.48% F1-score, 55.68% IoU, and 154 FPS for inference speed. Moreover, the accuracy for 133 aberrant EGD was increased to 96.24%. Regardless of speed, BlazeNeo with our strategies and modifications achieved the highest F1-score and IoU of 93.4% and 58.30%, respectively.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

In the future, transformer-based models can be investigated, but they must meet a real-time inference criterion regarding the inference speed greater than 100 frames per second (FPS). The existing demand for tracking GIM treatment has not yet been completed, and we are eager to investigate how our model can be utilized to assist professionals. To accomplish this, an approximation of the GIM size is required for each observation, with the intention that the size will decrease as a result of therapy. With the benefit of our model, segmentation can also be used to predict the GIM surface area. Such a possibility permits specialists to keep records of their treatment and provide patients with a more effective cure, thereby reducing the GC mortality rate.

CHAPTER VII

APPENDICES

7.1. A single model with a classifier

For a single model with a classifier (MobileNetV3s), the result is different depending on the model base architecture. BiSeNet performs better on a single model with an auxiliary head see Table 18; meanwhile, the best multi-color technique For STDC2-Seg50 model is the single model with a classifier. The model achieved 57.34% IoU and 91.35% F1-score, outperforming the single model with an auxiliary head by 4.81% and 1.29%, respectively. STDC2-Seg50 is a descendant of BiSeNet, which upgraded its backbone from ResNet-18 to STDC2 and replaced the spatial path with a skip connection. The ability to classify imaging modes (NBI and WLE) and segmentation GIM is excessive for a single path of STDC2-Seg50 compared to the dual path of ResNet-18 (see Figures 11 and 8, respectively). Thus, the single model with a classifier is reasonable for STDC2-Seg50. The effect of the multi-color techniques is also applied to BlazeNeo model as STDC2-Seg50; the single model with a classifier is the winner.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

7.1. An experiment of label smoothing with an edge cut and without.

Table 17 presents an effect of different label smoothing: with and without edge cut. The experiment was set on STDC2-Seg50 model combined with location-wise hard negative sampling and full jigsaw. Using label smoothing with an edge cut can leverage IoU score by approximately 4%; thus, label smoothing with an edge cut is preferable for STDC2-Seg50 model.

Table 17. *The effect of different label smoothing on the testing set was evaluated using STDC2-Seg50 model as the backbone. Boldface refers to the winner.*

Method	Acc	Sen	Spec	F1	PPV	NPV	IoU	Error
without an edge cut	**94.98**	88.17	**97.78**	91.11	**94.25**	95.25	53.67	**0.1667**
with an edge cut	94.48	**90.86**	96.19	**91.35**	91.85	**95.71**	**57.35**	0.2866

Table 18. *Performance comparison between a single model with an auxiliary head and a single model with an NBI&WLE classifier (MobileNetV3s). There are three parts in the Table 12: (1) BiSeNet model, (2) STDC2-Seg50 model, and (3) BlazeNeo model. Boldface refers to the winner for each model.*

Method	Acc	Sen	Spec	F1	PPV	NPV	IoU	Error	FPS
BiSeNet model									
with an auxiliary head	**94.31**	**91.40**	95.69	**91.15**	90.91	**95.93**	**54.66**	0.5256	**173**
with an NBI&WLE classifier (MobileNetV3s)	93.97	90.86	95.43	90.62	90.37	95.67	54.61	0.3706	135
STDC2-Seg50 model									
with an auxiliary head	94.35	87.63	97.12	90.06	92.61	95.01	52.53	0.3889	**153**
with an NBI&WLE classifier (MobileNetV3s)	**94.98**	**90.86**	**96.67**	**91.35**	**91.85**	**96.25**	**57.34**	**0.2512**	121
BlazeNeo model									
with an auxiliary head	95.29	89.25	97.78	**91.71**	94.32	95.66	54.16	0.2750	**93**
with an NBI&WLE classifier (MobileNetV3s)	**96.23**	**91.40**	**98.23**	93.41	**95.50**	**96.51**	**58.30**	**0.2253**	80

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

REFERENCES

1. Sung, H., et al., *Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries.* CA Cancer J Clin, 2021. **71**(3): p. 209-249.
2. Allemani, C., et al., *Global surveillance of trends in cancer survival 2000-14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries.* Lancet, 2018. **391**(10125): p. 1023-1075.
3. Take, I., et al., *Progress with each passing day: role of endoscopy in early gastric cancer.* Translational Gastrointestinal Cancer, 2015. **4**(6): p. 423-428.
4. Rokkas, T., M.I. Filipe, and G.E. Sladen, *Detection of an increased incidence of early gastric cancer in patients with intestinal metaplasia type III who are closely followed up.* Gut, 1991. **32**(10): p. 1110-3.
5. Panteris, V., et al., *Diagnostic capabilities of high-definition white light endoscopy for the diagnosis of gastric intestinal metaplasia and correlation with histologic and clinical data.* Eur J Gastroenterol Hepatol, 2014. **26**(6): p. 594-601.
6. Capelle, L.G., et al., *Narrow band imaging for the detection of gastric intestinal metaplasia and dysplasia during surveillance endoscopy.* Dig Dis Sci, 2010. **55**(12): p. 3442-8.
7. Pimentel-Nunes, P., et al., *A multicenter prospective study of the real-time use of narrow-band imaging in the diagnosis of premalignant gastric conditions and lesions.* Endoscopy, 2016. **48**(8): p. 723-30.
8. Huang, R.J., et al., *Diagnosis and Management of Gastric Intestinal Metaplasia: Current Status and Future Directions.* Gut Liver, 2019. **13**(6): p. 596-603.
9. Kanesaka, T., et al., *Computer-aided diagnosis for identifying and delineating early gastric cancers in magnifying narrow-band imaging.* Gastrointest Endosc, 2018. **87**(5): p. 1339-1344.
10. de Souza, L.A., Jr., et al., *A survey on Barrett's esophagus analysis using*

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

machine learning. Comput Biol Med, 2018. **96**: p. 203-213.

11. Baraldi, A. and F. Panniggiani, *An investigation of the textural characteristics associated with gray level cooccurrence matrix statistical parameters.* IEEE Transactions on Geoscience and Remote Sensing, 1995. **33**(2): p. 293-304.
12. Wichakam, I., et al. *Real-Time Polyps Segmentation for Colonoscopy Video Frames Using Compressed Fully Convolutional Network*. in *MultiMedia Modeling*. 2018. Cham: Springer International Publishing.
13. Sun, M., et al., *Accurate Gastric Cancer Segmentation in Digital Pathology Images Using Deformable Convolution and Multi-Scale Embedding Networks.* IEEE Access, 2019. **7**: p. 75530-75541.
14. Li, Y., et al. *GT-Net: A Deep Learning Network for Gastric Tumor Diagnosis*. in *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*. 2018.
15. Yu, T., et al., *Multi-label recognition of cancer-related lesions with clinical priors on white-light endoscopy.* Computers in Biology and Medicine, 2022. **143**: p. 105255.
16. Li, H., et al., *A multi-feature fusion method for image recognition of gastrointestinal metaplasia (GIM).* Biomedical Signal Processing and Control, 2021. **69**: p. 102909.
17. Lin, N., et al., *Simultaneous Recognition of Atrophic Gastritis and Intestinal Metaplasia on White Light Endoscopic Images Based on Convolutional Neural Networks: A Multicenter Study.* Clin Transl Gastroenterol, 2021. **12**(8): p. e00385.
18. Lai, Q., et al., *Multi-scale Multi-instance Multi-feature Joint Learning Broad Network (M3JLBN) for gastric intestinal metaplasia subtype classification.* Knowledge-Based Systems, 2022. **249**: p. 108960.
19. Wong, P.K., et al., *Broad learning system stacking with multi-scale attention for the diagnosis of gastric intestinal metaplasia.* Biomedical Signal Processing and Control, 2022. **73**: p. 103476.
20. Yan, T., et al., *Intelligent diagnosis of gastric intestinal metaplasia based on convolutional neural network and limited number of endoscopic images.* Computers in Biology and Medicine, 2020. **126**: p. 104026.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

21. Xu, M., et al., *Artificial intelligence in the diagnosis of gastric precancerous conditions by image-enhanced endoscopy: a multicenter, diagnostic study (with video).* Gastrointest Endosc, 2021. **94**(3): p. 540-548.e4.

22. Liu, X., et al., *Hue-texture-embedded region-based model for magnifying endoscopy with narrow-band imaging image segmentation based on visual features.* Comput Methods Programs Biomed, 2017. **145**: p. 53-66.

23. Wang, C., et al. *Localizing and Identifying Intestinal Metaplasia Based on Deep Learning in Oesophagoscope*. in *2019 8th International Symposium on Next Generation Electronics (ISNE)*. 2019.

24. Du, W., et al., *Automatic Early Gastric Cancer Segmentation in Gastroscopic Images Based on ResUnet*, in *2021 8th International Conference on Biomedical and Bioinformatics Engineering*. 2022, Association for Computing Machinery. p. 13–19 , numpages = 7.

25. Qiu, K., et al. *Research on ME- NBI Gastric Lesion Recognition System based on Improved UNet Structure*. in *2022 3rd International Conference on Electronic Communication and Artificial Intelligence (IWECAI)*. 2022.

26. Siripoppohn, V., et al. *Fast segmentation convolutional neural network with edge-guided path for real-time diagnosis of gastric intestinal metaplasia*. in *2021 25th International Computer Science and Engineering Conference (ICSEC)*. 2021.

27. Siripoppohn, V., et al., *Real-time semantic segmentation of gastric intestinal metaplasia using a deep learning approach.* Clin Endosc, 2022. **55**(3): p. 390-400.

28. Pornvoraphat, P., et al., *Real-time gastric intestinal metaplasia diagnosis tailored for bias and noisy-labeled data with multiple endoscopic imaging.* Computers in Biology and Medicine, 2023. **154**: p. 106582.

29. Research, M.F.f.M.E.a. *Upper endoscopy*. Available from: https://www.mayoclinic.org/tests-procedures/endoscopy/about/pac-20395197.

30. Gono, K., *Narrow Band Imaging: Technology Basis and Research and Development History.* Clin Endosc, 2015. **48**(6): p. 476-80.

31. Ronneberger, O., P. Fischer, and T. Brox. *U-Net: Convolutional Networks for*

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Biomedical Image Segmentation. 2015. Cham: Springer International Publishing.

32. Ekman, M., *Learning Deep Learning: Theory and Practice of Neural Networks, Computer Vision, NLP, and Transformers using TensorFlow*. 2021: Addison-Wesley Professional.

33. Yu, C., et al. *BiSeNet: Bilateral Segmentation Network for Real-Time Semantic Segmentation*. 2018. Cham: Springer International Publishing.

34. Chollet, F. *Xception: Deep Learning with Depthwise Separable Convolutions*. in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

35. Tan, M. and Q. Le, *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*, in *Proceedings of the 36th International Conference on Machine Learning*, C. Kamalika and S. Ruslan, Editors. 2019, PMLR: Proceedings of Machine Learning Research. p. 6105--6114.

36. Yu, C., et al., *BiSeNet V2: Bilateral Network with Guided Aggregation for Real-Time Semantic Segmentation.* International Journal of Computer Vision, 2021. **129**(11): p. 3051-3068.

37. Howard, A., et al., *MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications.* 2017.

38. Fan, M., et al., *Rethinking BiSeNet For Real-time Semantic Segmentation*. 2021. 9711-9720.

39. Hong, Y.a.P., Huihui and Sun, Weichao and Jia, Yisong, *Deep Dual-resolution Networks for Real-time and Accurate Semantic Segmentation of Road Scenes*. 2021: arXiv.

40. An, N.S., et al., *BlazeNeo: Blazing Fast Polyp Segmentation and Neoplasm Detection.* IEEE Access, 2022. **10**: p. 43669-43684.

41. Li, H., et al., *A multi-feature fusion method for image recognition of gastrointestinal metaplasia (GIM).* Biomedical Signal Processing and Control, 2021. **69**.

42. Yuan, Y., X. Chen, and J. Wang. *Object-Contextual Representations for Semantic Segmentation*. 2020. Cham: Springer International Publishing.

43. Robinson, J.a.C., Ching-Yao and Sra, Suvrit and Jegelka, Stefanie, *Contrastive*

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Learning with Hard Negative Samples. 2020: arXiv.

44. Januszewicz, W. and M.F. Kaminski, *Quality indicators in diagnostic upper gastrointestinal endoscopy.* Therapeutic Advances in Gastroenterology, 2020. **13**: p. 1756284820916693.

45. Cho, T.S., S. Avidan, and W.T. Freeman. *A probabilistic image jigsaw puzzle solver*. in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2010.

46. Frederick, M.W. and W.V.M. John. *Efficient algorithm for Gaussian blur using finite-state machines*. in *Proc.SPIE, Machine Vision Systems for Inspection and Metrology VII*. 1998. SPIE.

47. Howard, A., et al., *Searching for MobileNetV3*. 2019. 1314-1324.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

VITA

NAME	Passin Pornvoraphat
DATE OF BIRTH	16 April 1996
PLACE OF BIRTH	Bangkok, Thailand
INSTITUTIONS ATTENDED	Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University
HOME ADDRESS	188 Lat Phrao 48 Alley, Lane 8, Samsen Nok, Huai Khwang, Bangkok 10310
PUBLICATION	Pornvoraphat, P., et al., Real-time gastric intestinal metaplasia diagnosis tailored for bias and noisy-labeled data with multiple endoscopic imaging. Computers in Biology and Medicine, 2023. 154: p. 106582. Pornvoraphat, P., et al., Lift Analysis of Solar Floating under Aerodynamic Load. The 32nd Conference of Thailand's Mechanical Engineering Network, 2018.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY