

Chulalongkorn University

Chula Digital Collections

Chulalongkorn University Theses and Dissertations (Chula ETD)

2019

Semantic segmentation on remotely sensed images using deep convolutional encoder-decoder neural network

Teerapong Panboonyuen

Faculty of Engineering

Follow this and additional works at: <https://digital.car.chula.ac.th/chulaetd>

 Part of the [Computer Engineering Commons](#), and the [Computer Sciences Commons](#)

Recommended Citation

Panboonyuen, Teerapong, "Semantic segmentation on remotely sensed images using deep convolutional encoder-decoder neural network" (2019). *Chulalongkorn University Theses and Dissertations (Chula ETD)*. 8534.

<https://digital.car.chula.ac.th/chulaetd/8534>

This Thesis is brought to you for free and open access by Chula Digital Collections. It has been accepted for inclusion in Chulalongkorn University Theses and Dissertations (Chula ETD) by an authorized administrator of Chula Digital Collections. For more information, please contact ChulaDC@car.chula.ac.th.

Semantic Segmentation on Remotely Sensed Images Using Deep Convolutional
Encoder-Decoder Neural Network



A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in Computer Engineering

Department of Computer Engineering

FACULTY OF ENGINEERING

Chulalongkorn University

Academic Year 2019

Copyright of Chulalongkorn University

การแยกส่วนภาพทางความหมายบนภาพถ่ายระยะไกลโดยใช้โครงข่ายประสาทแบบคอนโวลูชันอล
เชิงลึกแบบเอ็นโค้ดเดอร์-ดีโค้ดเดอร์



วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรดุษฎีบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2562
ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

Thesis Title	Semantic Segmentation on Remotely Sensed Images Using Deep Convolutional Encoder-Decoder Neural Network
By	Mr. Teerapong Panboonyuen
Field of Study	Computer Engineering
Thesis Advisor	Assistant Professor PEERAPON VATEEKUL, Ph.D.
Thesis Co Advisor	Assistant Professor Kulsawasd Jitkajornwanich, Ph.D.

Accepted by the FACULTY OF ENGINEERING, Chulalongkorn University in
Partial Fulfillment of the Requirement for the Doctor of Philosophy

..... Dean of the FACULTY OF
ENGINEERING

(Professor SUPOT TEACHAVORASINSKUN, Ph.D.)

DISSERTATION COMMITTEE

..... Chairman

(Professor BOONSERM KIJSIRIKUL, Ph.D.)

..... Thesis Advisor

(Assistant Professor PEERAPON VATEEKUL, Ph.D.)

..... Thesis Co-Advisor

(Assistant Professor Kulsawasd Jitkajornwanich, Ph.D.)

..... Examiner

(Associate Professor THANARAT CHALIDABHONGSE, Ph.D.)

..... Examiner

(Ekapol Chuangsuwanich, Ph.D.)

..... External Examiner

(Siam Lawawirojwong, Ph.D.)

..... External Examiner

(Panu Srestasathiern, Ph.D.)

ธีรพงศ์ ปานบุญยืน : การแยกส่วนภาพทางความหมายบนภาพถ่ายระยะไกลโดยใช้โครงข่ายประสาทแบบคอนโวลูชันอลเซิงลิกแบบอัลเอนโค้ดเดอร์-ดิโค้ดเดอร์. (Semantic Segmentation on Remotely Sensed Images Using Deep Convolutional Encoder-Decoder Neural Network) อ.ที่ปรึกษาหลัก : ผศ. ดร.พีรพล เวทีกุล, อ.ที่ปรึกษาร่วม : ผศ. ดร.กุลสวัสดิ์ จิตชารawanich

การจำแนกทางความหมายออกจากภาพถ่ายทางอากาศและภาพถ่ายดาวเทียมเป็นหนึ่งในงานพื้นฐานของการรับรู้ระยะไกล มีบทบาทสำคัญในการใช้งานเชิงโปรแกรมประยุกต์ เช่น การวางแผนการเกษตร การปรับปรุงแผนที่ การเพิ่มประสิทธิภาพเส้นทาง และการนำทางไปยังที่ต่างๆ โดยเดลที่เป็นมาตรฐานที่ดีที่สุดคือเน็ตเวิร์กคอนโวลูชันเชิงลึกแบบอัลเอนโค้ดเดอร์-ดิโค้ดเดอร์ อย่างไรก็ตามค่าความถูกต้องยังถูกจำกัด เพราะสถาปัตยกรรมนี้ไม่ได้ถูกออกแบบมาเพื่อจับคุณสมบัติระดับต่ำบนภาพถ่ายระยะไกล เช่น วัตถุแม่น้ำ วัตถุที่เป็นพืชพันธุ์ต่างๆ อีกทั้งปริมาณของข้อมูลฝึกสอนในงานประเภทนี้ยังมีไม่เพียงพอ วิทยานิพนธ์นี้จึงเสนอวิธีการปรับปรุงและออกแบบสถาปัตยกรรม การจำแนกทางความหมายในห้าขั้นตอน เริ่มจากขั้นตอนแรกได้เสนอเน็ตเวิร์กคอนโวลูชันเชิงลึก ปรับแต่งแบบโกลบอลซึ่งเป็นหนึ่งในเน็ตเวิร์กคอนโวลูชันสมัยใหม่ ขั้นตอนที่สองประยุกต์ใช้ความสนใจตามช่องเพื่อเลือกตัวกรองที่ดี ขั้นตอนที่สามนำการเรียนรู้แบบถ่ายโอนเฉพาะโดเมนมาใช้ สำหรับแก้ปัญหาการขาดแคลนของข้อมูลฝึกสอน ขั้นตอนที่สี่ทำการเพิ่มคุณสมบัติพิเศษเข้าไป ที่เน็ตเวิร์กเพื่อจับคุณสมบัติภาพในระดับล่าง และขั้นตอนสุดท้ายประยุกต์ใช้คอนโวลูชันแบบเดฟไวส์อตรัส ชุดข้อมูลที่ใช้ทำการทดลองมีทั้งหมด 3 ชุด ประกอบด้วย สองชุดข้อมูลที่จัดเก็บเองมา จากราบเทียม Landsat-8 และหนึ่งชุดข้อมูลมาจากชุดข้อมูลมาตรฐานจากรายการการแข่งขัน ISPRS Vaihingen โดยมีหนึ่งวิธีการมาตราฐานที่เป็นงานมาตรฐานที่ดีที่สุด คือ Deep Convolutional Encoder-Decoder (DCED) ผลการทดลองด้วยวิธีที่นำเสนอแสดงให้เห็นว่า ประสิทธิภาพที่ได้ดีกว่าวิธีการมาตรฐาน

6071467821 : MAJOR COMPUTER ENGINEERING

KEYWORD: Convolutional neural network, Deep Learning, Remote Sensing, Feature fusion, Transfer learning, Semantic Segmentation, Computer Vision, Landsat-8, Aerial Image, Satellite Image

Teerapong Panboonyuen : Semantic Segmentation on Remotely Sensed Images Using Deep Convolutional Encoder-Decoder Neural Network.

Advisor: Asst. Prof. PEERAPON VATEEKUL, Ph.D. Co-advisor: Asst. Prof. Kulsawasd Jitkajornwanich, Ph.D.

One of the fundamental tasks in remote sensing is the semantic segmentation of the aerial and satellite images. It plays a vital role in applications, such as agriculture planning, map updates, route optimization, and navigation. The state-of-the-art model is the Deep Convolutional Encoder-Decoder (DCED). However, the accuracy is still limited since the architecture is not designed for recovering low-level features, e.g., river, low vegetation on remotely sensed images, and the training data in this domain are deficient. In this dissertation, we aim to propose the semantic segmentation architecture in five aspects, designed explicitly for the remotely sensed field. First, we propose applying a modern Convolutional Neural Network (CNN) called a Global Convolutional Network (GCN). Second, “channel attention” is presented to select the most discriminative filters (features). Third, “domain-specific transfer learning” is introduced to alleviate the scarcity issue. Fourth, “Feature Fusion (FF)” is added to our network to capture low-level features. Finally, “Depthwise Atrous Convolution (DA)” is introduced to refine the extracted features. The experiment was conducted on three data sets: two private corpora from Landsat-8 satellite and one public benchmark from the “ISPRS Vaihingen” challenge. The results showed that our proposed architectures outperformed the baseline model on any remote sensing imagery.

Field of Study: Computer Engineering Student's Signature

Academic Year: 2019 Advisor's Signature

Co-advisor's Signature

ACKNOWLEDGEMENTS

I would like to thank many persons who assisted me along with my journey for writing this dissertation. I feel felicitous enough not to have experienced the hackneyed solitary of Ph.D. research life.

Firstly, I would like to thank my advisor, Professor Peerapon Vateekul, for his explicit guidance and encouragement, including all of my co-authors, Professor Kulsawasd Jitkajornwanich, Dr. Siam Lawawirojwong, and Dr. Panu Srestasathiern as well and I especially learned a great deal from working with them. I benefited not only from his profound insights and knowledge but also from his patience. I am also grateful to Professor Boonserm Kijsirikul, Professor Thanarat Chalidabhongse, and Professor Ekapol Chuangsawanich for serving on my supervisory committees and for providing valuable feedback throughout. I also want to thank all the current and former members of the Data Mining Group (DataMind) and Machine Intelligence and Knowledge Discovery Lab (MIND Lab) for contributing to a truly remarkable, fun research environment, and for many interesting discussions.

Next, I appreciate and thanks to the scholarship from the 100th Anniversary Chulalongkorn University Fund for the Doctoral Scholarship and the 90th Anniversary Chulalongkorn University Fund (Ratchadaphiseksomphot Endowment Fund). I much acknowledge the Geo-Informatics and Space Technology Development Agency (GISTDA), Thailand. I would like to thank the staff from the GISTDA (Thanwarat Anan, Bussakon Satta, and Suwalak Nakya) for providing the remote sensing corpora used in this study.

Finally, I would also like to thank my grandfather and grandmother for their never-ending support and for raising me to value education. Most importantly, I would like to thank my family—Panboonyuen—and my friends, who kept the conversation academic for only what was necessary.

Teerapong Panboonyuen

TABLE OF CONTENTS

	Page
ABSTRACT (THAI)	iii
ABSTRACT (ENGLISH)	iv
ACKNOWLEDGEMENTS	v
TABLE OF CONTENTS.....	vi
LIST OF TABLES	x
LIST OF FIGURES	xi
CHAPTER I.....	1
INTRODUCTION.....	1
1.1 Aims and Objectives.....	6
1.2 The scope of work.....	6
1.3 Publication (selected only ISI and Scopus database since 2018 to 2020)	7
CHAPTER II.....	8
BACKGROUND	8
2.1 Neural Networks	8
2.2. Convolutional neural network (ConvNets or CNNs).....	9
2.2.1 Convolution Layer	10
2.2.2 Pooling Layer	11
2.2.3 Deconvolution Layer.....	11
2.2.3 Un-pooling Layer.....	12
2.3. Loss Function	13
2.3.1 Quadratic Cost Function.....	13

2.3.2 Cross Entropy Cost Function	13
2.3.3 Exponential Cost Function.....	14
2.4. Gradient Descent Optimization	14
2.4.1 Adagrad	14
2.4.1 Adadelta	15
2.4.1 RMSprop (Root Mean Square Propagation)	15
2.4.1 Adam (Adaptive Moment Estimation).....	15
CHAPTER III.....	16
RELATED WORK.....	16
3.1. Deep Learning Concepts for Semantic Segmentation.....	16
3.2. Modern Deep Learning Architectures for Semantic Segmentation.....	18
3.3. Modern Techniques of Deep Learning	19
3.4. Deep Learning for Semantic Segmentation on Aerial and Satellite Images	22
3.5. How to Design the Modern Deep Learning for Segmentation Task on the Landsat-8 Satellite and the ISPRS Vaihingen Challenge Corpora	23
CHAPTER IV	26
CONCEPTS AND RESEARCH METHODOLOGY	26
4.1. Data Preprocessing.....	27
4.2. A Global Convolutional Network (GCN) with Variations of Backbones	28
4.3. The Channel Attention Block (A)	29
4.4. Domain-Specific Transfer Learning (TL).....	31
4.5. Feature Fusion Concept.....	32
4.6. Depthwise Atrous Convolution (DA).....	35
CHAPTER V	38

EXPERIMENTAL RESULTS	38
5.1. Landsat-8w3c Dataset	38
5.2. Landsat-8w5c Dataset	39
5.3. ISPRS Vaihingen Dataset.....	41
5.3. Evaluation	43
5.4. Experimental Setup	43
5.5. Results of the Landsat-8w5c Corpus with Discussion.....	46
5.5.1. The Effect of an Enhanced GCN on the Landsat-8w5c Corpus	47
5.5.2. The Effect of Using Channel Attention on the Landsat-8w5c Corpus.....	48
5.5.3. The Effect of Using Domain-Specific Transfer Learning on LS8w5c Corpus	48
5.5.4. The Effect of Using Feature Fusion on Landsat-8w5c Corpus.....	48
5.5.5. The Effect of Using Depthwise Atrous Convolution on Landsat-8w5c Corpus.....	49
5.6. Results of the Landsat-8w3c Corpus with Discussion.....	59
5.6.1. The Effect of an Enhanced GCN on the Landsat-8w3c Corpus	59
5.6.2. The Effect of Using Channel Attention on the Landsat-8w3c Corpus.....	59
5.6.3. The Effect of Using Domain-Specific Transfer Learning on LS8w3c Corpus	60
5.6.4. The Effect of Using Feature Fusion on Landsat-8w3c Corpus.....	60
5.6.5. The Effect of Using Depthwise Atrous Convolution on Landsat-8w3c Corpus.....	61
5.7. Results of the ISPRS Vaihingen Challenge Corpus with Discussion.....	72
5.7.1. Effect of the Enhanced GCN on the ISPRS Vaihingen Corpus.....	72
5.7.2. Effect of Using Channel Attention on ISPRS Vaihingen Corpus.....	73

5.7.3. The Effect of Using Domain-Specific Transfer Learning on the ISPRS Corpus.....	74
5.7.4. The Effect of Using Feature Fusion on ISPRS Vaihingen Corpus	74
5.7.5. The Effect of Using Depthwise Atrous Convolution on on ISPRS Vaihingen Corpus.....	74
CHAPTER VI	83
CONCLUSIONS	83
REFERENCES	85
VITA.....	93



LIST OF TABLES

	Page
Table 1. Performance comparison of existing models on CamVid corpus	20
Table 2. Performance comparison of existing models on Cityscapes corpus	21
Table 3. Performance comparison of existing models on remote sensing corpus	23
Table 4 Abbreviations on our Landsat-8 corpora.....	38
Table 5. Abbreviations on our proposed deep learning methods.....	45
Table 6. Results of the testing data of the Landsat-8 (Nan) corpus between baseline and 7 variations of proposed techniques in terms of precision, recall, and F1-score.	46
Table 7. Results of the testing data of Landsat-8 corpus (Nan) between each class with our proposed techniques in terms of accuracy.....	47
Table 8. Results of the testing data of the Isan corpus between baseline and 6 variations of proposed techniques in terms of precision, recall, and F1-score	62
Table 9. Results of the testing data of Landsat-8 (Isan) corpus between each class with our proposed techniques in terms of accuracy	62
Table 10. Results of the testing data of the ISPRS 2D semantic labeling challenge corpus between the baseline and five variations of our proposed techniques in terms of precision, recall, and F1-score	72
Table 11. Results of the testing data of ISPRS Vaihingen Challenge corpus between each class with our proposed techniques in terms of accuracy.....	73

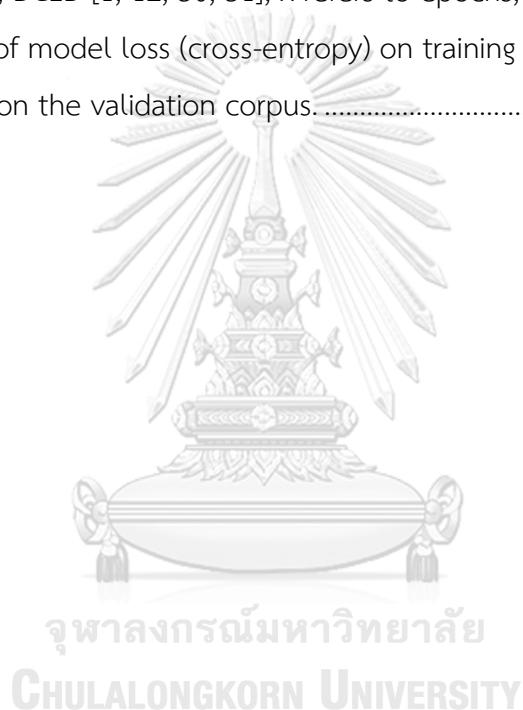
LIST OF FIGURES

	Page
Figure 1. Sample of remote sensing image. Input image (left), and target image (right).	3
Figure 2. False positive and false negative samples on the aerial images. It cannot predict the low-level features object, such as low vegetation.	4
Figure 3. False positive and false negative samples on the satellite images (Landsat-8w5c). It cannot predict the low-level features object, such as river.	5
Figure 4. False positive and false negative samples on the satellite images (Landsat-8w5c). It cannot predict the low-level features object, such as pineapple.	5
Figure 5. Overview of neural networks [40] (a) The building block of deep neural networks and (b) Example of a feed-forward multilayer neural network	8
Figure 6. An example of deep CNNs architecture. [41]	9
Figure 7. An example of convolution of image	10
Figure 8. An example of max-pooling with a (3x3) kernel	11
Figure 9. Illustration of deconvolution operations [43]	12
Figure 10. Illustration of Un-pooling operations [43]	12
Figure 11. An overview of the whole original Global Convolutional architecture [33] in (A) The details of Global Convolutional Network (GCN) and Boundary Refinement (BR) block are represented in (B) and (C), consecutively	19
Figure 12. Illustration of each type of deep learning architectures [14, 62] (a) presents the VGG style deep structure (b) indicates the U-shape structure and (c) demonstrates the context-path style	25
Figure 13. An overview of our proposed network (I)	26

Figure 14. An overview of our full proposed network (II). The GCN152-TL-FF-DA [74]: an enhanced GCN architecture with feature fusion and depthwise atrous convolution.	27
.....	
Figure 15. An overview of the whole backbone pipeline in (left) the main backbone with varying by ResNet50, ResNet101, and ResNet152; (right) the major drivers of our main classification network (composed of a global convolutional network (GCN) and a boundary refinement (BR) block [33])	29
.....	
Figure 16. Components of the channel attention block. The red lines represent the down sample operators, respectively. The red line cannot change the size of feature maps. It is only a path for information passing	30
.....	
Figure 17. The domain-specific transfer learning strategy reuses pre-trained weights of models between two datasets—very high (ISPRS) and medium (Landsat-8; LS-8) resolution images	32
.....	
Figure 18. The framework of our feature fusion strategy	33
.....	
Figure 19. The Depthwise Atrous Convolution (DA) module in the proposed parallel pyramid method for improving feature fusion	37
.....	
Figure 20. Sample satellite images from Isan (Northeastern Thailand), a zone in Thailand (left), and corresponding ground truth (right). The label of this data set includes three categories: corn (yellow), pineapple (green), and rubber tree (red)	40
.....	
Figure 21. Sample satellite images from Nan, a province in Thailand (left), and corresponding ground truth (right). The label of medium resolution dataset includes five categories: agriculture (yellow), forest (green), miscellaneous (brown), urban (red), and water (blue)	40
.....	
Figure 22. Overview of the ISPRS 2D Vaihingen Labeling corpus. There are 33 tiles. Numbers in the figure refer to the individual tile flag	42
.....	
Figure 23. The sample input tile from Figure 7 (left) and corresponding ground truth (right). The label of the Vaihingen Challenge includes six categories: impervious	

surface (imp surf, white), building (blue), low vegetation (low veg, cyan), tree (green), car (yellow), and clutter/background (red)	42
Figure 24. Testing sample inputs and output satellite images on Landsat-8w5c in the Nan province in Thailand, where rows refer to different images. (a) Original input image. (b) Target map (ground truth). (c) Output of Encoder–Decoder (Baseline). (d) Output of GCN152-TL-A. (e) Output of GCN152-TL-A-FF. and (f) Output of GCN152-TL-A-FF-DA. The label of medium resolution dataset includes five categories: Agriculture (yellow), Forest (green), Miscellaneous (Misc, brown), Urban (red) and Water (blue) .	57
Figure 25. Graph (learning curves) on Landsat-8w5c data set of the proposed approach, “GCN152-TL-A-FF-DA”; x refers to epochs, and y refers to different measures (a) Plot of model loss (cross-entropy) on training and validation corpora; (b) performance plot on the validation corpus.	58
Figure 26. Graph (learning curves) on Landsat-8w5c data set of the baseline approach, DCED [1, 12, 30, 31]; x refers to epochs, and y refers to different measures (a) Plot of model loss (cross-entropy) on training and validation corpora; (b) performance plot on the validation corpus.	58
Figure 27. Testing sample input and output satellite images on Landsat-8 in Isan (Northeastern) in Thailand, where rows refer to different images. (a) Original input image. (b) Target map (ground truth). (c) Output of DCED (Baseline). (d) Output of GCN152-TL-A. (e) Output of GCN152-TL-A-FF. and (f) Output of GCN152-TL-A-FF-DA.	70
Figure 28. Graph (learning curves) on Landsat-8w3c data set of the proposed approach, “GCN152-TL-A-FF-DA”; x refers to epochs, and y refers to different measures (a) Plot of model loss (cross-entropy) on training and validation corpora; (b) performance plot on the validation corpus.	71
Figure 29. Graph (learning curves) on Landsat-8w3c data set of the baseline approach, DCED [1, 12, 30, 31]; x refers to epochs, and y refers to different measures (a) Plot of model loss (cross-entropy) on training and validation corpora; (b) performance plot on the validation corpus.	71

Figure 30. Comparisons between “GCN152-TL-A-FF-DA” and beyond baseline methods on the ISPRS Vaihingen (Stuttgart) challenge corpus testing set	79
Figure 31. Graph (learning curves) on ISPRS Vaihingen Challenge corpus data set of the proposed approach, “GCN152-TL-A-FF-DA”; x refers to epochs, and y refers to different measures (a) Plot of model loss (cross-entropy) on training and validation corpora; (b) performance plot on the validation corpus.....	80
Figure 32. Graph (learning curves) on ISPRS Vaihingen Challenge corpus of the baseline approach, DCED [1, 12, 30, 31]; x refers to epochs, and y refers to different measures (a) Plot of model loss (cross-entropy) on training and validation corpora; (b) performance plot on the validation corpus.....	81



CHAPTER I

INTRODUCTION

Semantic segmentation on remote sensing images (as shown in Figure 1) is a long-standing research problem. It has been implemented in many applications in various domains, e.g., urban planning, map updates, route optimization, and navigation [1-8] allowing us to better understand the domain's images and create important real-world applications. Natural objects such as roads, water, forests, urban, and agriculture fields regions are operated in various tasks to create imperative remotely sensed applications. The target of this problem is to assign each pixel to a given object category. As a result of research articles [1, 2], this task is very challenging, for rare class and small object class, such as low vegetation and water objects.

A deep convolutional neural network (CNN, CNNs, or ConvNet) is a well-known method for automatic feature learning. It can automatically learn features at different levels and abstractions from raw images by multiple hierarchical stacking convolution and pooling layers. To accomplish such a challenging task, features at different levels are required. Specifically, abstract high-level features are more suitable for the recognition of confusing manmade objects, while the labeling of finely structured objects could benefit from detailed low-level features [1, 2, 6]. Therefore, different numbers of layers will affect the performance of deep learning models [9-11].

A deep convolutional encoder-decoder (DCED) architecture, one of the most efficient newly developed neural networks, has been proposed for object segmentation and given good performance in the experiments tested on CamVid¹,

¹ <http://mi.eng.cam.ac.uk/research/projects/VideoRec/CamVid/>

Cityscapes², and PASCAL VOC 2012³ data—a well-known benchmark corpus for image segmentation research and has inherently encoded different levels of a feature. Instinctively, some methods integrate them to refine the final prediction. This branch of methods mainly considers how to recover the reduced spatial information caused by consecutive pooling operator or convolution with stride. There are many encoder-decoder networks [9, 10, 12-29] for semantic segmentation. For example, SegNet [12, 30, 31] utilizes the saved pool indices to recover the reduced spatial information. This network uses a VGG-style encoder-decoder, where the up-sampling in the decoder is done using transposed convolutions. U-net [32] uses the skip connection and consists of a contracting path and an expansive path, which gives it the u-shaped architecture. PSPNet [18] utilizes the capability of global context information by different-region based context aggregation. However, this type of architecture ignores the global context. Besides, most methods of this type are just summed up the features of adjacent stages without consideration of their diverse representation. This leads to some inconsistent results.

In the past few years, the modern CNNs have been extensively proposed including Global Convolutional Network (GCN) [33] in which the large kernel and effective receptive field play an important role in performing classification and localization tasks simultaneously. GCN is proposed to address the classification and localization issues for semantic segmentation and to suggest a residual-based boundary refinement for further refining object boundaries. However, this type of architecture ignores the global context such as weights of the features in each stage. Furthermore, most methods of this type are just summed up the features of adjacent stages without considering their diverse representations. This leads to some inconsistent results that suffer from accuracy performance. The primary challenge of this remote sensing task is a lack of training data. It has become a motivation of this work. Nevertheless, the state of the

² <https://www.cityscapes-dataset.com/>

³ <http://host.robots.ox.ac.uk/pascal/VOC/voc2012/>

art work in this field still disregards the local context, such as low-level features in each stage. Moreover, most feature fusion methods are just a summation of the features from adjacent stages and they do not consider the representations of diversity (critical for the performance of the CNN). This leads to unpredictable results that suffer from measuring the performance such as the F1 score. This, in fact, is the inspiration for this work as well.

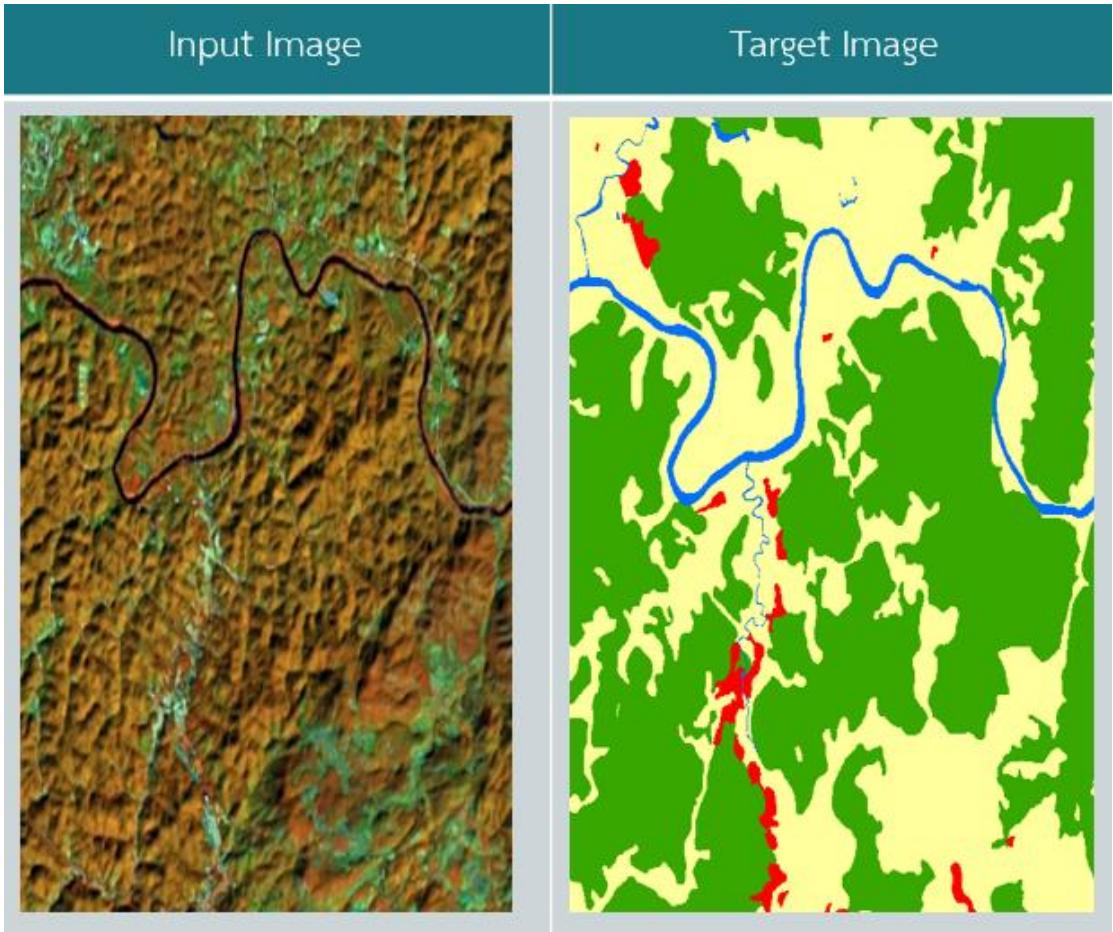


Figure 1. Sample of remote sensing image. Input image (left), and target image (right).

Although the current baseline methods [1, 12, 30, 31, 34] have achieved significant breakthroughs in semantic labeling on remote sensing corpora, it is still laborious to manually label the MR images in river and pineapple areas and the VHR images in low vegetation and car areas. The two reasons are as follows: (i) previous approaches are

less efficient to recover low-level features for accurate labeling, and (ii) they ignore the low-level features learned by the backbone network's shallow layers with long-span connections, which is caused by semantic gaps in different-level contexts and features. We show the problem of false-positive and false-negative samples on remote sensing, as shown in Figures 2, 3, and 4, consecutively.

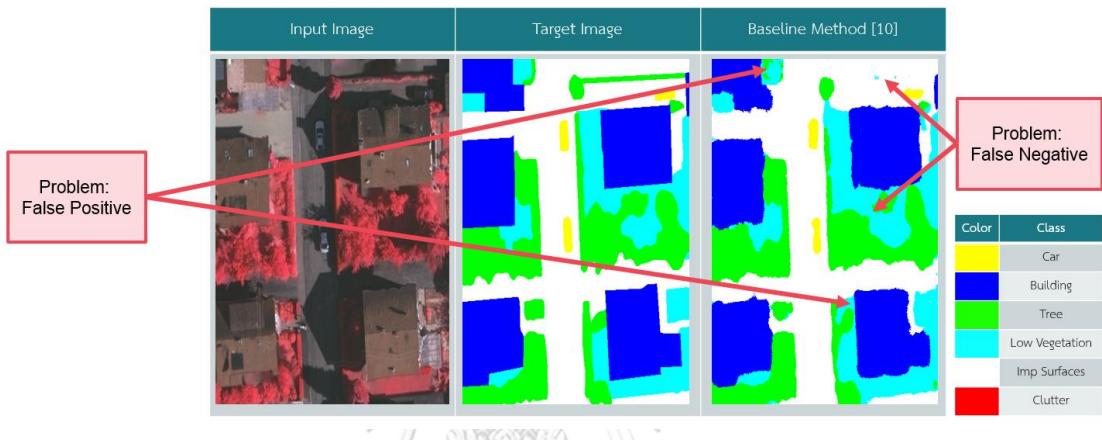


Figure 2. False positive and false negative samples on the aerial images. It cannot predict the low-level features object, such as low vegetation.

In this dissertation, we present a novel global convolutional network for segmenting multi-objects from aerial and satellite images. To this end, it is focused on five aspects: (i) varying backbones using ResNet50, ResNet101, and ResNet152, (ii) applying a "channel attention block" [14, 35, 36] to assign weights for feature maps in each stage of the backbone architecture, (iii) employing "domain-specific transfer learning" [37-39] to relieve scarcity, (iv) feature fusion concept is proposed to fuse the different in layer of feature representation for capture mostly rich detail information such as low-level class (for example, car and river class), and the last, (v) depthwise atrous convolution (DA) is proposed to bridge the semantic gap and implement durable multi-level feature aggregation to extract complementary information from very shallow features. Experiments were conducted using satellite imagery (from the

Landsat-8 satellite), which was provided by a government organization in Thailand, and using well-known aerial imagery from the ISPRS Vaihingen Challenge corpus [1, 2, 29], which is publicly available. The results showed that our method outperforms the baseline including deep convolutional encoder–decoder (DCED) in terms of F1-score.

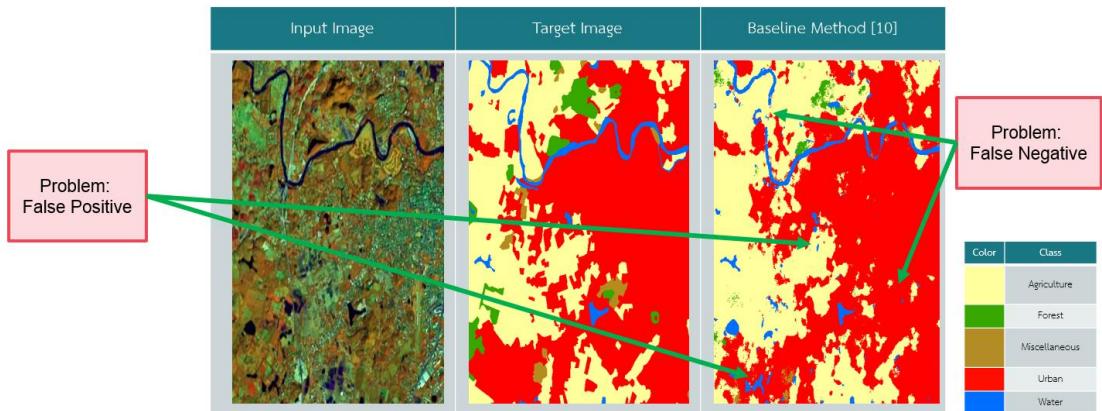


Figure 3. False positive and false negative samples on the satellite images (Landsat-8w5c). It cannot predict the low-level features object, such as river.

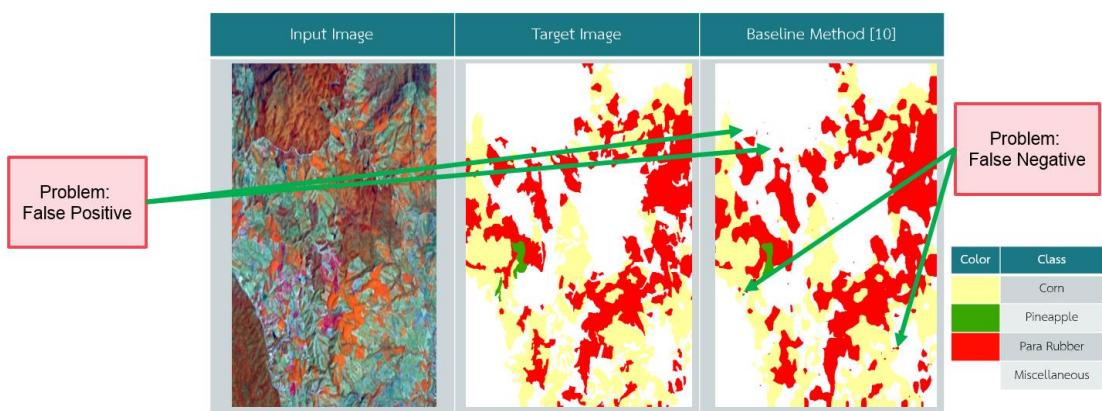


Figure 4. False positive and false negative samples on the satellite images (Landsat-8w5c). It cannot predict the low-level features object, such as pineapple.

The rest of the dissertation is organized as follows:

- Chapter 2 presents a brief background of existing work on applying deep learning to remote sensing images.
- Chapter 3 discusses the related work of deep learning concepts for semantic segmentation on the challenge and remote sensing corpora.
- Chapter 4 presents concepts and research methodology.
- Chapter 5 shows our whole experiments.
- Chapter 6 concludes our most important Findings and offers a discussion of the most promising directions for improving our full proposed method.

1.1 Aims and Objectives

1. To propose a new deep learning architecture to segment multi-objects from aerial and satellite images
2. To explore the effectiveness of the proposed new deep learning techniques for semantic segmentation particularly on remote sensing corpora

1.2 The scope of work

1. Evaluate the proposed new deep learning on ISPRS Vaihingen corpus (a city district of Stuttgart, Germany) and GISTDA⁴ corpora (Nan province and Isan zone corpora) with DCED baseline model
 - a. GISTDA Nan province corpora have five classes: agriculture, forest, miscellaneous, urban, and water
 - b. GISTDA Isan zone corpora have three classes: corn, pineapple, and rubber tree
2. Evaluate the proposed deep learning on reliable measurements such as Precision, Recall, and F1-score

⁴Geo-Informatics and Space Technology Development Agency (Public Organization)

<https://www.gistda.or.th/main/en>

1.3 Publication (selected only ISI and Scopus database since 2018 to 2020)

- **Panboonyuen, T.**; Jitkajornwanich, K.; Lawawirojwong, S.; Srestasathiern, P.; Vateekul, P. "Semantic Segmentation on Remotely Sensed Images Using an Enhanced Global Convolutional Network with Channel Attention and Domain Specific Transfer Learning ". *Remote Sens.* 2019, 11, 83.
 - Remote Sensing, ISI Journal, Q1 with Tier 1, Rank 1
 - Impact Factor = 3.406
 - <https://www.mdpi.com/2072-4292/11/1/83>
- **Panboonyuen, T.**; Jitkajornwanich, K.; Lawawirojwong, S.; Srestasathiern, P.; Vateekul, P. Semantic Labeling in Remote Sensing Corpora Using Feature Fusion-Based Enhanced Global Convolutional Network with High-Resolution Representations and Depthwise Atrous Convolution. *Remote Sens.* 2020, 12, 1233.
 - Remote Sensing, ISI Journal, Q1 with Tier 1, Rank 1
 - Impact Factor = 4.118
 - <https://www.mdpi.com/2072-4292/12/8/1233>
- Wichakam, I., **Panboonyuen, T.**, Udomcharoenchaikit, C., & Vateekul, P. (2018, February). Real-Time Polyps Segmentation for Colonoscopy Video Frames Using Compressed Fully Convolutional Network. In International Conference on Multimedia Modeling (pp. 393-404). Springer, Cham.
- Chantharaj, S., Pornrathanapong, K., Chitsinpchayakun, P., **Panboonyuen, T.**, Vateekul, P., Lawawirojwong, S., ... & Jitkajornwanich, K. (2018, July). Semantic Segmentation on Medium-Resolution Satellite Images Using Deep Convolutional Networks with Remote Sensing Derived Indices. In 2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE) (pp. 1-6). IEEE.
- **My Web of Science ResearcherID: AAO-4985-2020**
(<https://publons.com/researcher/AAO-4985-2020/>)
- **MY ORCID: Connecting Research and Researchers**
(<https://orcid.org/0000-0001-8464-4476>)

CHAPTER II

BACKGROUND

In this chapter, the background knowledge related to the dissertation is presented.

2.1 Neural Networks

It is a network of neurons that are used to process information. To create these, it is constructed from 3 types of layers: the first is Input layer that initial data for the neural network. Next, the second is hidden layers—intermediate layers between input and output layer and place where all the computation is done. Last, output layer is to produce the result for given inputs.

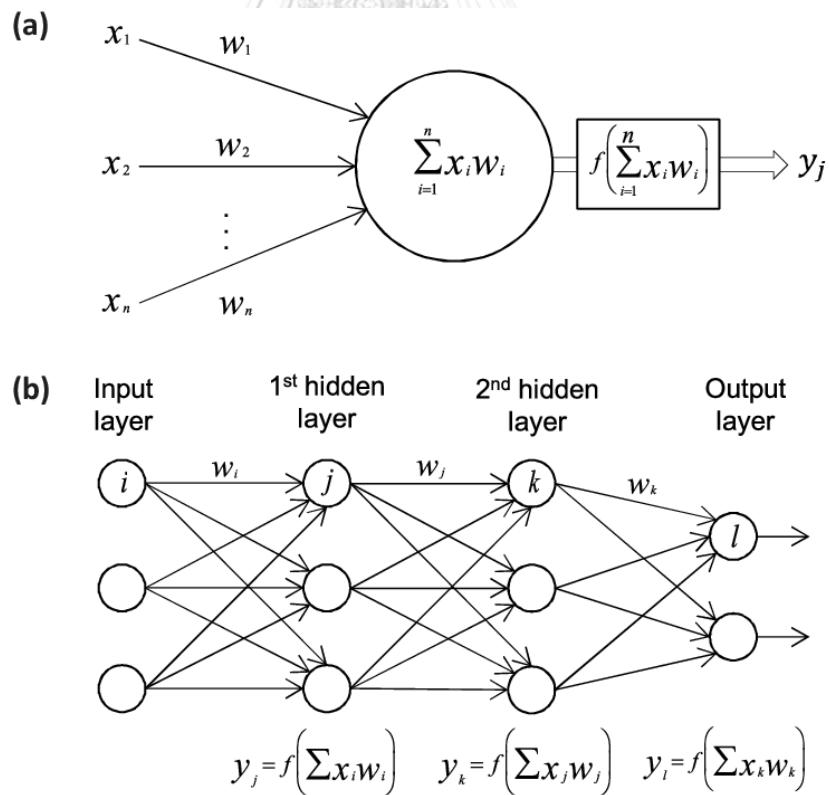


Figure 5. Overview of neural networks [40] (a) The building block of deep neural networks and (b) Example of a feed-forward multilayer neural network

Figure 5a, each input x_i has an associated weight w_i . The sum of all weighted inputs, $x_i w_i$, is then passed through a nonlinear activation function f , to transform the pre-activation level of the neuron to an output y_j . Figure 5b, a feedforward multilayer neural network with binary classes, in which the nodes in one layer are connected to all neurons in the fully connected network. The information is propagated through the network up to the output layer, where the softmax function yields the probability of a given observation belonging to each class.

2.2. Convolutional neural network (ConvNets or CNNs)

CNNs are one of the main components for doing semantic segmentation. Technically, deep learning CNN models⁵ to train and test, each input image will pass it through a series of convolution layers (Conv in Figure 6) with filters (kernels), pooling layers (Pool in Figure 6), fully connected layers (FC in Figure 6) and apply softmax function (Output in Figure 6) to classify an object with probabilistic values between 0 and 1. The below figure is a complete flow of CNN to process an input image and classifies the objects based on values.

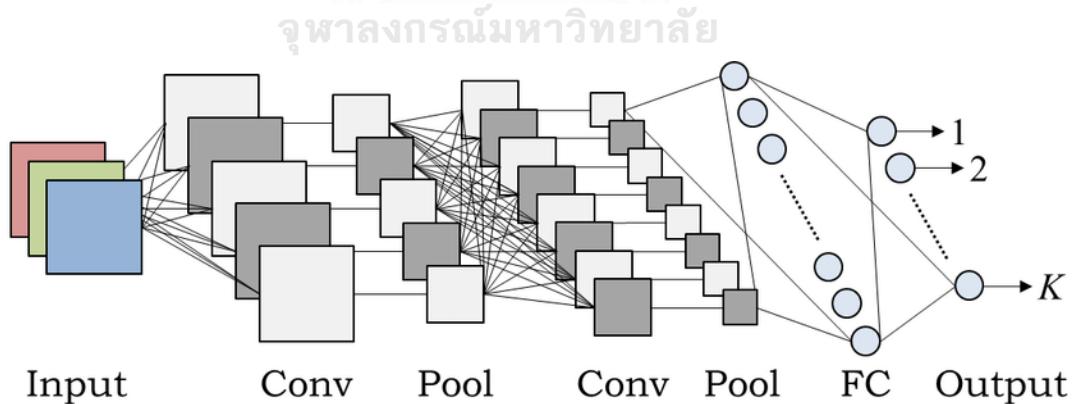


Figure 6. An example of deep CNNs architecture. [41]

⁵ <http://cs231n.stanford.edu/>

2.2.1 Convolution Layer

It is the first layer to extract features from an input image (Figure 7). It preserves the relationship between pixels by learning image features using small squares of input data. It is a mathematical operation that takes two inputs such as image matrix and a filter or kernel. An image therefore has size $h \times w \times d$ where color channel depth $d = 3$. Convolutional layers are essential layers in CNNs [7, 8, 40-42], producing feature maps from input images or lower level feature maps.

Equations 1 show the relationship between output size O and input size of an image I after convolution with stride s and kernel K . Furthermore, the feature map size decreases as the number of convolutional layers increases. Row output size O_x and column output size O_y of convolutional layers are determined as follows:

$$\begin{aligned} O_x &= \frac{I_x - K_x}{s} + 1, \\ O_y &= \frac{I_y - K_y}{s} + 1 \end{aligned} \quad (1)$$

For example, we have an image of size $(32 \times 32 \times 3)$, by a kernel of size $(3 \times 3 \times 3)$ and a stride $s = 1$ result in an activation map of size $(30 \times 30 \times 1)$. Using additional n kernels, the activation map becomes $(30 \times 30 \times n)$. Therefore, further kernels will increase the depth of the convolutional layer output.

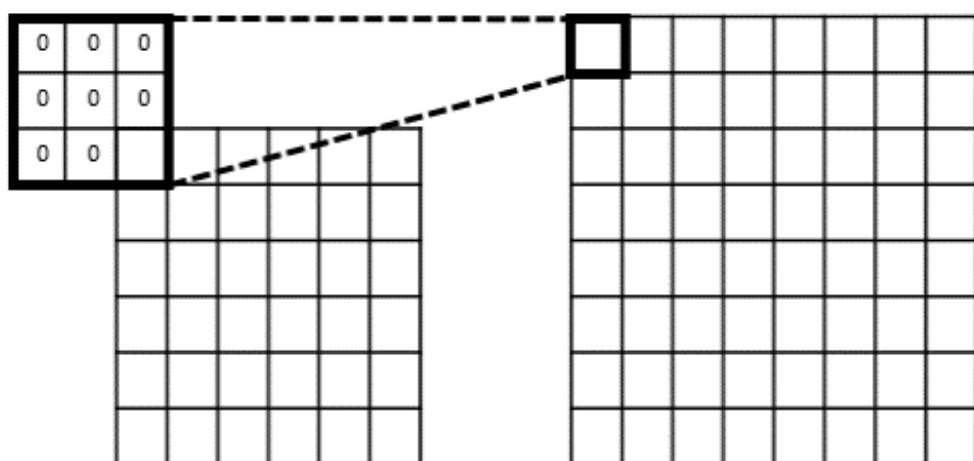


Figure 7. An example of convolution of image

2.2.2 Pooling Layer

This layers are also known as subsampling or down-sampling layers [7, 8, 40-42]. A commonly used pooling method is max-pooling (Figure 8). The down-sampled output is produced by taking the maximum input value within the filter, resulting in an output of decreased size. There are many methods which are commonly used in this layer of CNNs, such as average pooling and L2-norm pooling. A pooling Layer operates independently on every depth slice of the input and resizes it spatially, using the max operation. A pooling layer has a kernel and a stride of similar length.

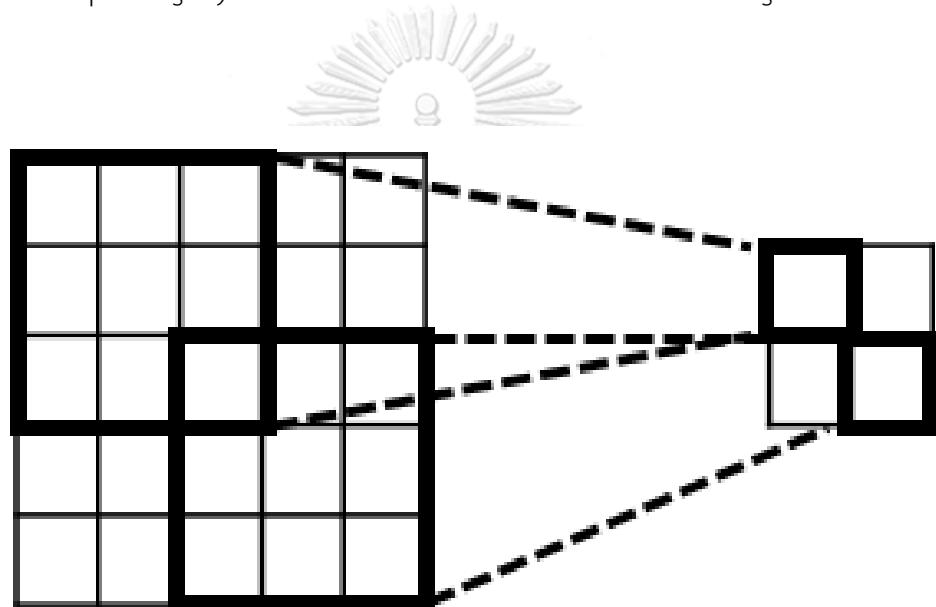


Figure 8. An example of max-pooling with a (3x3) kernel

2.2.3 Deconvolution Layer

It also called transposed convolutions or fractionally stridden convolutions (Figure 9), is a layer which can obtain a dense map from down-sampled and course input. This layers make the sparse activations obtained by un-pooling through convolution-like operations with multiple learned filters.

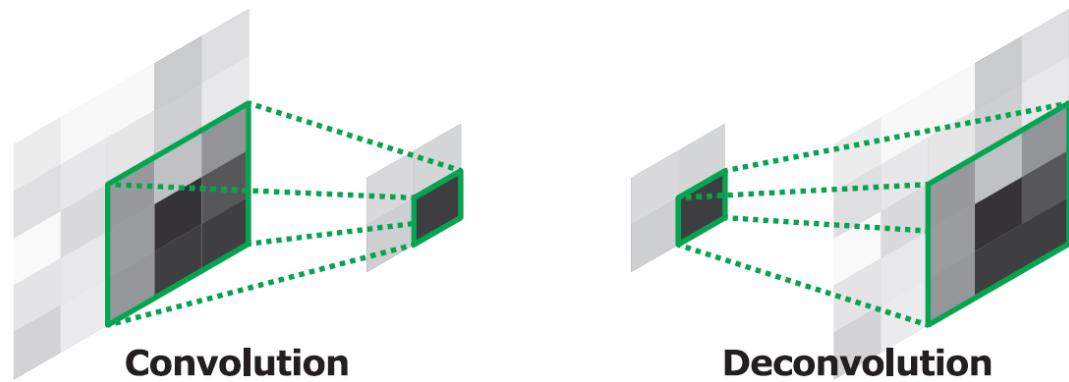


Figure 9. Illustration of deconvolution operations [43]

2.2.3 Un-pooling Layer

Refers to the original idea [32, 40, 42-44], the un-pooling operation uses these switches to place the reconstructions from the layer above into appropriate locations, preserving the structure of the stimulus. Shown as Figure 10 for an illustration of Un-pooling operations.

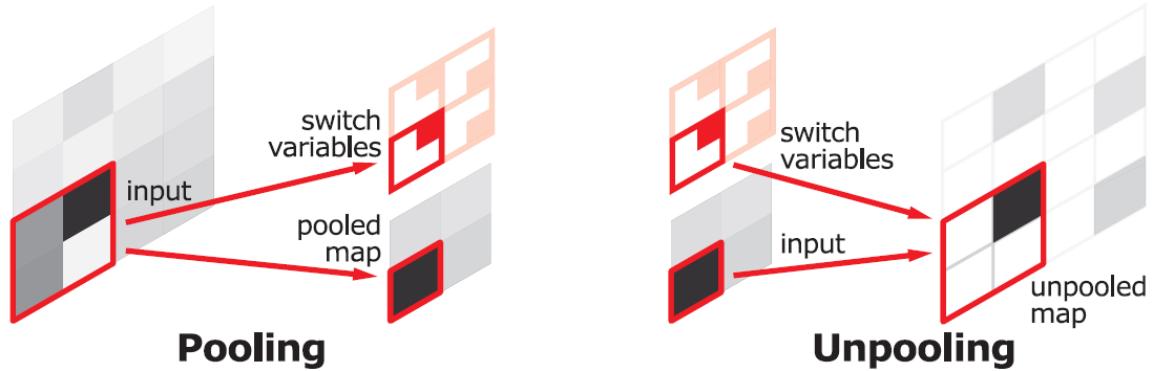


Figure 10. Illustration of Un-pooling operations [43]

2.3. Loss Function

The value of the loss function [8, 32, 40, 42-44] is L ; It represents the difference between the training image after it has propagated through the network and desired annotated output image. Two inferences are made about this loss function. (i) it should be able to define the loss function as the average over the loss functions for individual training data set, as the training often is carried out in batches. The loss function is evaluated and average at the end of each batch, then the weights are updated. Next, (ii) the loss function should be able to be defined as a function of the network outputs. Below a brief overview is given of some widely used loss functions, where x_i are the neuron outputs and \hat{x}_i are the desired outputs.

2.3.1 Quadratic Cost Function

This function (equation 2) is also known as Mean Squared Error (MSE) cost function. It is one of the simplest cost functions.

$$L = \frac{1}{N} \sum_{i=1}^N (x_i - \hat{x}_i)^2 \quad (2)$$

2.3.2 Cross Entropy Cost Function

The cross entropy cost function⁶ (equation 3) is commonly used in convolutional network applications.

$$L = \frac{1}{N} \sum_{i=1}^N (\hat{x}_i \ln(x_i) + (1 - \hat{x}_i) \ln(1 - x_i)) \quad (3)$$

⁶ <https://medium.com/datadriveninvestor/overview-of-different-optimizers-for-neural-networks-e0ed119440c3>

2.3.3 Exponential Cost Function

The exponential cost function (equation 4) requires an additional parameter τ .

$$L = \frac{1}{N} \tau \exp \frac{1}{\tau} \sum_{i=1}^N (x_i - \hat{x}_i)^2 \quad (4)$$

2.4. Gradient Descent Optimization

There are several variants of optimizer available. Determining the appropriate learning rate, or step size, often is a complex problem. Applying too high learning rates causes suboptimal performance, too low learning rates cause slow convergence. Learning rate schedule is used as an extension of the optimizer algorithm to increase performance

2.4.1 Adagrad

It adapts the updates to the slope of the error function. The algorithm adapts the learning rate to the parameters, so that size of the updates for each parameter depends on its importance. The Adagrad algorithm gives larger updates for infrequent parameters and smaller updates for frequent parameters, the update rule is given in equation 5.

CHULALONGKORN UNIVERSITY

$$\begin{aligned} g_{t,i} &= \nabla_{\theta} L(\theta_i), \\ \theta_{t+1,i} &= \theta_{t,i} - \frac{n}{\sqrt{G_{t,ii} + \epsilon}} \cdot g_{t,i} \end{aligned} \quad (5)$$

2.4.1 Adadelta

It is an extended version of the original from Adagrad which reduces the problem of the decreasing learning rate. It restricts the range of accumulated squared gradients to a certain fixed size.

2.4.1 RMSprop (Root Mean Square Propagation)

This function is also an adaptive learning rate method that tackles the problem of the accumulation of squared gradients in Adagrad. RMSprop⁷ divides the learning rate by an exponentially decaying average of squared gradients. It is an unpublished algorithm by G. Hinton.

It also tries to dampen the oscillations, but in a different way than momentum. It also takes away the need to adjust learning rate, and does it automatically. More so, RMSProp chooses a different learning rate for each parameter.

2.4.1 Adam (Adaptive Moment Estimation)

Adam (equation 6) can be seen as a combination of RMSprop and Stochastic Gradient Descent with momentum. It is an adaptive learning rate method. It computes individual learning rates for different parameters. It also determines an adaptive learning rate for each parameter and keeps an exponentially decaying average of past gradients.

$$\theta_{t+1} = \theta_t - \frac{n}{\sqrt{\hat{v}_t} + \epsilon} \cdot \hat{m}_t \quad (6)$$

⁷ <https://blog.paperspace.com/intro-to-optimization-momentum-rmsprop-adam/>

CHAPTER III

RELATED WORK

Deep learning is one of the fast-growing fields in machine learning which has been successfully applied to computer vision tasks and has been successfully applied for remotely sensed data analysis, notably land cover mapping on urban areas [4-6, 8, 41, 45] and has increasingly become a promising tool for accelerating the image recognition process with high accuracy [7, 9-11, 13-23, 25-28, 32, 41, 43, 44, 46-65]. It is really a fast-growing field, and new architectures appear every few days. This chapter is divided into five sub-chapters: (i) Deep learning concepts for semantic segmentation. (ii) Modern deep learning architectures for semantic segmentation. (iii) Advanced techniques of deep learning. (vi) Deep learning for semantic segmentation on the remote sensing field. And (v) last, how to design modern deep learning.

3.1. Deep Learning Concepts for Semantic Segmentation

Semantic segmentation algorithms are often formulated to solve structured pixel-wise labeling problems based on a deep CNN. Noh et al. [43] proposed a novel semantic segmentation technique utilizing a deconvolutional neural network (DCNN) and the top layer from the DCNN adopted from VGG16 [26, 29, 35]. The DCNN structure is composed of up-sampling layers and deconvolution layers, describing pixel-wise class labels and predicting segmentation masks, respectively. Their proposed deep learning methods yield high performance in PASCAL VOC 2012 corpus, with the 72.5% accuracy in the best-case scenario (the highest accuracy—as of the time of the writing of this dissertation—compared to other methods that were trained without requiring additional or external data). Long et al. [44] proposed adapted contemporary classification networks incorporating Alex, VGG, and GoogLe networks into a fully CNN.

In this method, some of the pooling layers were skipped: Layer 3 (FCN-8s), Layer 4 (FCN-16s), and Layer 5 (FCN-32s). The skip architecture reduces the potential overfitting problem and has shown improvements in performance, ranging from 20% to 62.2% in the experiments tested on PASCAL VOC 2012 data. Ronneberger et al. [32] proposed U-Net, a DCNN for biomedical image segmentation. The architecture consists of a contracting path and a symmetric expanding path that captures context and consequently enables precise localization. The proposed network claimed to be capable of learning despite the limited number of training images and performed better than the prior best method (a sliding-window DCNN) on the ISBI challenge for segmentation of neuronal structures in electron microscopic stacks. Vijay Badrinarayanan [12, 30, 31] proposed a deep convolutional encoder-decoder network (DCED), called "SegNet" that consists of two main networks, encoder and decoder, and some outer layers. The two outer layers of the decoder network are responsible for feature extraction, the results of which are transmitted to the layer adjacent to the last layer of the decoder network. This layer is responsible for pixel-wise classification (determining which pixel belongs to which class). There is no fully connected layer in between feature extraction layers. In the up-sampling layer of the decoder, pool indices from the encoder are distributed to the decoder, where the kernel will be trained in each epoch (the training round) at the convolution layer. In the last layer (classification), softmax was used as a classifier for pixel-wise classification. The DCED is one of the deep learning models that exceeds the state of the art on many remote sensing corpus.

In this work, the DCED method was selected as our baseline since it is the most popular architecture used in various networks for semantic segmentation.

3.2. Modern Deep Learning Architectures for Semantic Segmentation

Recently, many approaches based on the DCED have achieved high performance on different benchmarks [12, 26, 30, 31, 43, 44] such as CamVid, PASCAL VOC 2012 and Cityscapes corpora. However, most of them still suffer from accuracy performance issues. Therefore, many works of modern deep learning architectures have been proposed, such as pyramid scene parsing network [18], the capability of global context information by different-region based context aggregation is applied through a pyramid pooling module together with the proposed pyramid scene parsing network. Liang-Chieh Chen and their friends from Google Inc. [13] propose DeepLabV3+ network which adds a decoder module on top of the regular DeepLabV3 model. Simon Jégo et al. [24] present down-sampling and up-sampling style encoder-decoder network. In addition, it concatenated skip connections from the encoder to the decoder and extend DenseNets to deal with the problem of semantic segmentation. Instance-aware semantic segmentation [48], which is slightly different from semantic segmentation. Instead of labeling all pixels, it focuses on the target objects and labels only pixels of those objects. FCIS [9] is based on techniques based on fully convolutional networks (FCNs). BiSeNet [14] use a spatial path with a small stride to preserve the spatial information and generate high-resolution features while having a parallel context path with a fast down-sampling strategy to obtain sufficient receptive field and design two specific modules: (i) Feature Fusion Module (FFM) and (ii) Attention Refinement Module (ARM), to further improve the accuracy. The mask R-CNN [27] was built around the FCN and is incorporated with a proposed joint formulation. Peng [33] presented the concept of large kernel matters to improve semantic segmentation with a global convolutional network (GCN) as shown in Figure 11. They proposed a GCN to address both the classification and localization issues for semantic segmentation. Large separable kernels were used to expand the receptive field, and a boundary refinement block was added to further improve localization performance near the boundaries. From the

CamVid⁸ corpus (created by Machine Intelligence Laboratory, Cambridge University), the GCN outperforms methods of all previous publications (all modern deep learning baselines) and has become the new state of the art. Therefore, the GCN was selected as our proposed method and as the main model of our work.

From Table 1 and Table 2, the GCN architecture was selected as our baseline since it is the winner architecture for semantic segmentation on CamVid corpus and Cityscapes corpus.

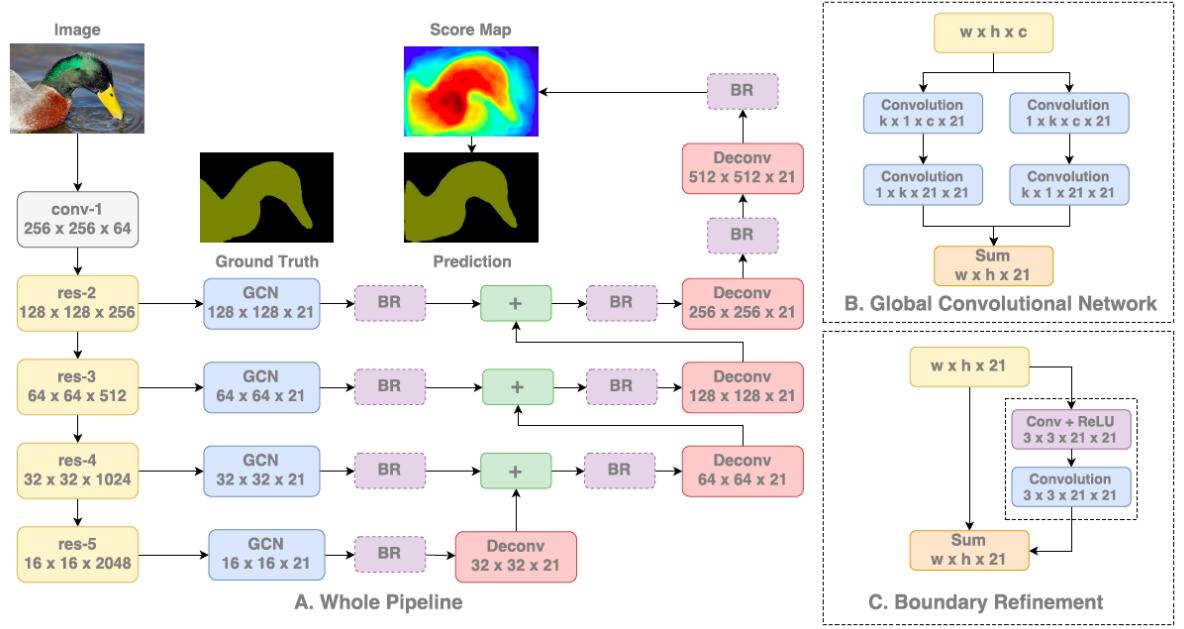


Figure 11. An overview of the whole original Global Convolutional architecture [33] in (A) The details of Global Convolutional Network (GCN) and Boundary Refinement (BR) block are represented in (B) and (C), consecutively

3.3. Modern Techniques of Deep Learning

Modern techniques of deep learning are important for the accuracy of a CNN. The most popular modern ideas used for semantic segmentation tasks, such as global

8 <http://mi.eng.cam.ac.uk/research/projects/VideoRec/CamVid/>

context, the attention module, and semantic boundary detection, have been used for boosting accuracy.

Global context [33] is a modern method that has proven the effectiveness of global average pooling in the semantic segmentation task. For example, PSPNet [18] and Deeplab v3 [11, 25, 58] respectively extend it to spatial pyramid pooling [18] and atrous spatial pyramid pooling [25], resulting in great performance at different benchmarks. However, to take advantage of the pyramid pooling module sufficiently, these two methods adopt the base feature network to downsample with atrous convolution eight times [25], which is time-consuming and memory-intensive.

Table 1. Performance comparison of existing models on CamVid corpus

Deep Learning Model	Precision	Recall	F1-Score
PSPNet [18]	0.74	0.74	0.74
DenseNet (Tiramisu) [24]	0.74	0.77	0.75
GCN [33]	0.85	0.87	0.86
DeepLabV3 [13]	0.72	0.63	0.67
BiseNet [14]	0.84	0.82	0.83

Attention Module [14, 35, 36]: Attention is helpful to focus on what we want. Recently, the attention module has increasingly become a powerful tool for deep neural networks. The method in [14, 35, 36] pays attention to different scale information. In this work, we utilize a channel attention block to select features, similar to learning a discriminative feature network [35]. Atrous convolution [11, 13, 17, 18, 66, 67], also known as multi-scale context aggregation, is proposed to regularly aggregate multi-scale contextual information devoid of losing resolution. In this paper, we use the technique of “Depthwise Atrous Convolution (DA)” [67] to extract complementary

information from very shallow features and enhance the deep features for improving feature fusion from our feature fusion step.

Table 2. Performance comparison of existing models on Cityscapes corpus

Method	mean-IoU
FCN 8s [44]	0.653
DPN [18]	0.591
CRFasRNN	0.625
Scale invariant CNN + CRF	0.663
Dilation10	0.671
DeepLabv2-CRF [11]	0.704
Adelaide	0.718
LRR-4x	0.716
Enocer Decoder [12, 30, 31]	0.754
GCN [33]	0.769



Refinement Residual Block [35]: The feature maps of each stage in the feature network all go through the refinement residual block. For our work, we use the boundary refinement block (BR) to be a concept of "refinement residual block" from [33]. The first component of the block is a 1×1 convolution layer. We use it to unify the number of channels to 21. Meanwhile, it can combine the information across all channels. Then the following is a basic residual block [58, 68], which can refine the feature map. Furthermore, this block can strengthen the recognition ability of each stage, inspired from the architecture of ResNet.

Feature Fusion Module [14, 69-73]: They fuse the features of the two paths are different in level of feature representation for accurate and recovering the rich information. So, to fuse these features they first concatenate the output features of Spatial Path and Context Path and then utilize the batch normalization [59] to balance the scales of the features. Final, they pool the concatenated feature to a feature vector and compute a weight vector, like SENet [36]. This weight vector can reweight the features, which amounts to feature selection and combination. Figure 14 shows the details of feature fusion design.

3.4. Deep Learning for Semantic Segmentation on Aerial and Satellite Images

Recently, many approaches based on the Encoder-Decoder have achieved high performance on remote sensing corpus (ISPRS Vaihingen challenge). Wang et al. [2] propose a gated convolutional neural network for the semantic segmentation in only high-resolution images, called gated segmentation network (GSN) by combining two feature maps (two paths are different in level of feature representation). Their proposed deep learning methods yield high performance in ISPRS Vaihingen corpus, with the 85.2% F1-score. Liu and their friends [1] propose Encoder-Decoder by focusing on three methods: (1) they present a novel deep encoder-decoder neural network (DCED based-ScasNet, multi-scale contexts aggregation) for distinguishing confusing manmade objects; (2) utilization of low-level features for fine structured objects refinement; (3) residual correction for more effective multi-feature fusion. Their proposed deep learning methods yield the highest performance (winner) in ISPRS Vaihingen corpus, with the 87.4% F1-score (Table 3).

On the other hand, there is still one of ISPRS corpus (very high-resolution images), called Potsdam corpus which like ISPRS Vaihingen corpus (just difference between a place that capturing aerial imagery. So, in this dissertation we will select and experiment only one data set for aerial images data set, which is ISPRS Vaihingen

corpus). For ISPRS Potsdam corpus, Diakogiannis et al. [68] introduce a novel Fully Convolutional Network (FCN) for semantic segmentation, called ResUNet, combines ideas distilled from computer vision applications of deep learning, and demonstrates competitive performance. They tested using the Potsdam data set made available through the ISPRS competition (ISPRS) yield the performance in ISPRS Potsdam corpus, with the 92.9% F1-score.

From Table 3, the novel deep encoder-decoder neural network (DCED based-ScasNet) architecture was selected as our baseline since it is the winner architecture for semantic segmentation on ISPRS Vaihingen corpus.

Table 3. Performance comparison of existing models on remote sensing corpus

Method	Imp surf	Building	Low veg	Tree	Car	F1-score
FCN-8s [44]	0.871	0.918	0.752	0.861	0.638	0.808
SegNet [30]	0.867	0.891	0.763	0.839	0.657	0.847
DeconvNet [43]	0.891	0.932	0.814	0.857	0.684	0.835
Deeplab [14]	0.892	0.945	0.749	0.875	0.798	0.852
DCED [1]	0.872	0.893	0.841	0.914	0.815	0.854

3.5. How to Design the Modern Deep Learning for Segmentation Task on the Landsat-8 Satellite and the ISPRS Vaihingen Challenge Corpora

Deep learning is commonly developed at a fixed resource cost and then scaled up in order to achieve better performance (F-score) when more resources are made available. For example, EfficientNets [62] can be scaled up from ResNet-50 (EfficientNets-B0) to EfficientNets-B(1-7) by increasing the number of layers, and recently, it can achieve 84.4% ImageNet top-1 accuracy by scaling up a baseline CNN

by a factor of four. The conventional practice for model scaling is to arbitrarily increase the CNN depth or width layer for training model and evaluate the performance.

Fundamentally, there are 3 main types (A, B, and C) (Figure 12) to design the new deep learning architecture for increasing both higher accuracy and better efficiency. It equally scales each dimension with a fixed set of scaling coefficients. The effectiveness of deep learning model scaling also relies weightily on the best baseline network (GCN network). Therefore, to further improve performance, we have also proposed the new deep learning network by performing follows 4 proposed methods for improving performance. The detail will be described in the next chapter.

For example in [17], an “ASPP” module is to design to capture context information of different receptive field by using Type B (U-Shape Style). PSPNet [18] applies a “PSP” module which contains several different scales of average pooling layers. Yang et al. [17] designs an “ASPP” module with a global average pooling to capture the global context of the image by also using Type B (U-Shape Style). Scale-adaptive convolutions [46] improves the neural network by a scale adaptive convolution layer to obtain an adaptive field context information by using Type A (VGG Style). DFN [16] adds the global pooling on the top of the U-shape architecture to encode the global context by using Type C (Context Path Style).

- **Type A:** VGG Style, It is conventional scaling that only increases the depth dimension of deep learning architecture
- **Type B:** U-Shape Style, It is conventional scaling that only increases the width dimension of deep learning architecture by fusing the hierarchical features of the backbone network
- **Type C:** Context Path Style, It is conventional scaling that both increases depth and width dimension of deep learning architecture

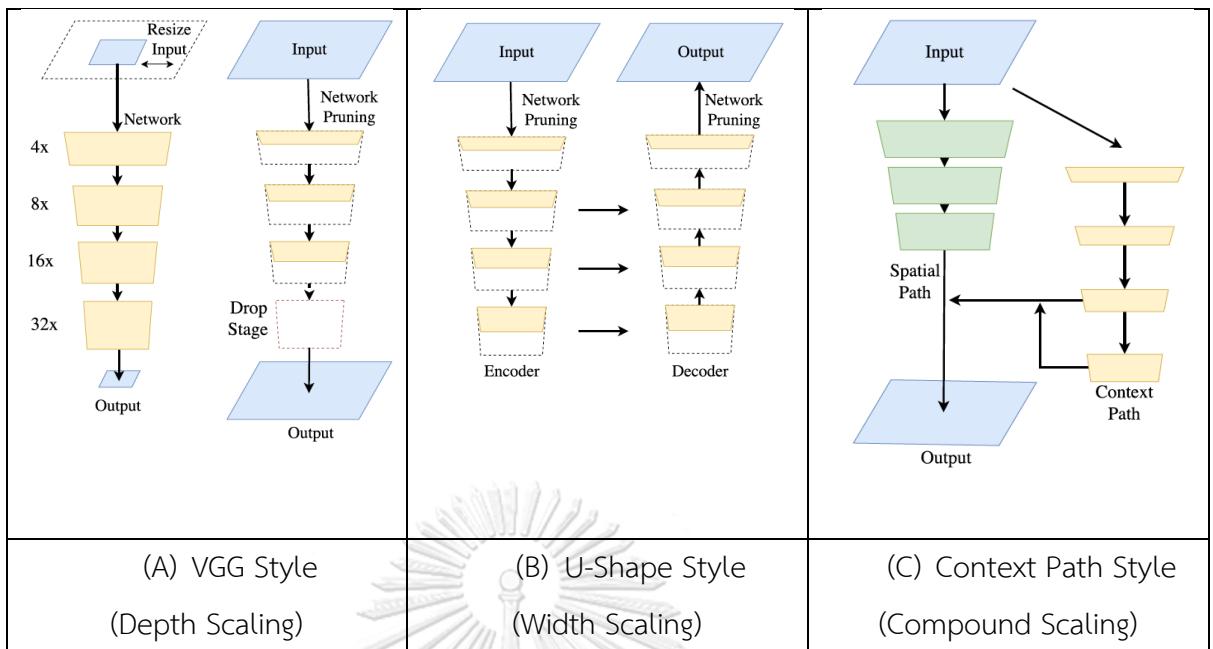


Figure 12. Illustration of each type of deep learning architectures⁹ [14, 62] (a) presents the VGG style deep structure (b) indicates the U-shape structure and (c) demonstrates the context-path style

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

⁹ <https://ai.googleblog.com/2019/05/efficientnet-improving-accuracy-and.html>

CHAPTER IV

CONCEPTS AND RESEARCH METHODOLOGY

In this chapter, the details of our proposed network are explained (shown in Figure 13). The network is based on the GCN with three aspects of improvements: (i) the modification of backbone architecture (shown in P1 in Figure 13), (ii) applying the channel attention block (shown in P2 in Figure 13), (iii) using the concept of domain-specific transfer learning (shown in P3 in Figure 13), (iv) proposed feature fusion module (shown in P4 in Figure 13 and Figure 14) to fuse the different in layer of feature representation for capture mostly rich detail information, and Last, (v) using the concept of “Depthwise Atrous Convolution” (shown in P5 in Figure 14). We called the full proposed method as "**Encoders Matter**".

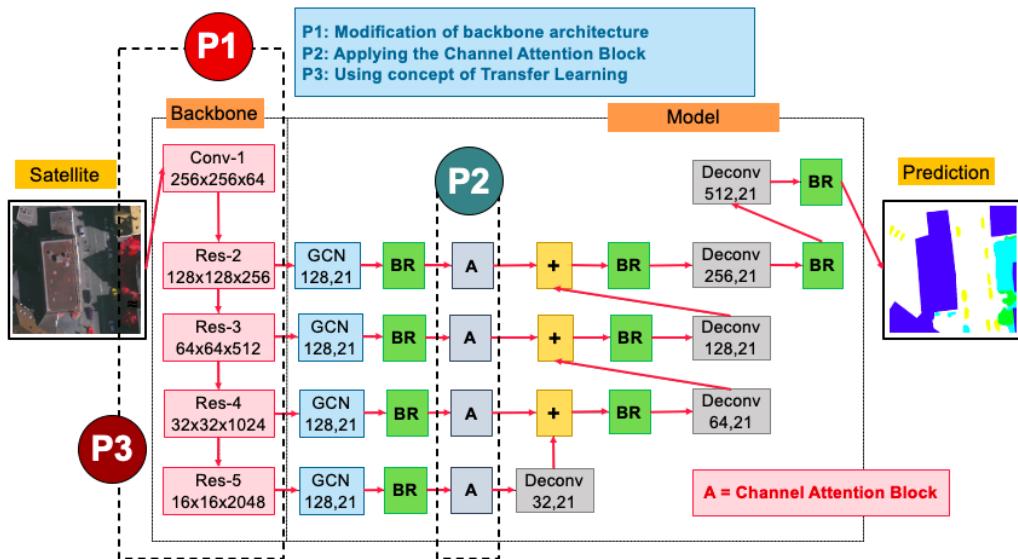


Figure 13. An overview of our proposed network (I)

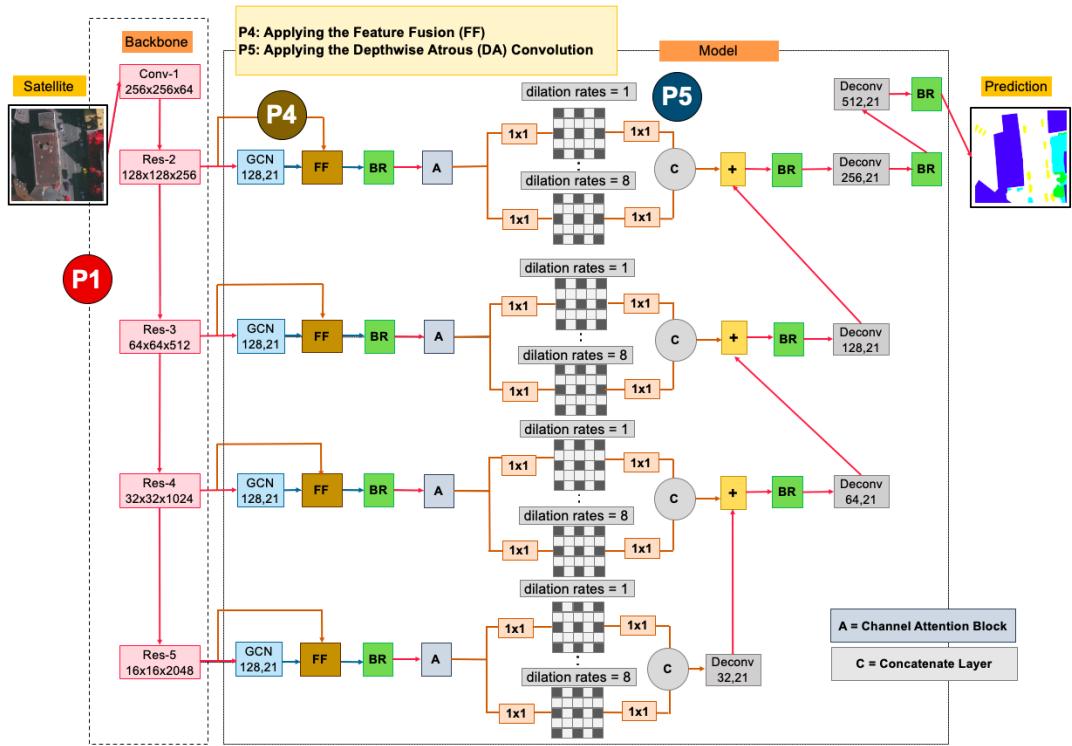


Figure 14. An overview of our full proposed network (II). **The GCN152-TL-FF-DA** [74]: an enhanced GCN architecture with feature fusion and depthwise atrous convolution.

4.1. Data Preprocessing

In this dissertation, there are two benchmark corpuses, including (i) the ISPRS Vaihingen Challenge corpus and (ii) the Landsat-8 dataset. They are comprised of very high and medium resolution images, respectively. More details of the datasets will be explained in Chapter 5.1, Chapter 5.2, and Chapter 5.3. Before a discussion of the model, it is worth explaining our data preprocessing procedure, since it is required when working with neural network and deep learning models. Thus, the mean subtraction is executed.

In addition, data augmentation is often required on more complex object recognition tasks. Therefore, a random horizontal flip is generated to increase the training data. For the ISPRS corpus, all images are standardized and cropped into 512x512 pixels with a resolution of 9 cm²/pixel. For the Landsat-8 corpus, each

image is also flipped horizontally and scaled to 512×512 with a resolution of 30 m²/pixel from the original images (16,800 × 15,800 pixels).

4.2. A Global Convolutional Network (GCN) with Variations of Backbones

GCN [33] as shown in Figure 11 is a modern architecture that surpasses the drawbacks of a traditional semantic segmentation network, such as deep convolutional encoder-decoder (DCED) networks. A traditional network usually cascades convolutional layers in order to generate sophisticated features; they can be considered as local features that are specialized for a specific task. However, it is not necessary to employ only specialized features; the general features are also important. Thus, a GCN overcomes this issue by introducing a multi-level architecture, where each level aims to capture a different resolution of features, so both local and global features are considered in the model.

As shown in Figure 11, there are two main blocks in the GCN: a localization block and a classification block. From the localization view in the left block, the structure is a stack of classical fully convolutional layers called "levels". Each level aims to construct features with different resolutions. From the classification view, there are two modules: the GCN and the boundary refinement (BR). For the GCN module, the kernel size of the convolutional structure should be as large as possible, which is motivated by the densely connected structure of the classification models. If the kernel size increases to the spatial size of the feature map (named the global convolution), the network will share the same benefits with the pure classification models. The BR module is added to further improve localization performance near the boundaries.

Although the GCN architecture has shown promising prediction performance, it is still possible to further improve by varying backbones using ResNet [58] with different numbers of layers as ResNet50, ResNet101, and ResNet152, as shown in Figure 15.

Additionally, the GCN is suggested to work on a large kernel size. In this dissertation, we set the large kernel size as 7 (this previous work [33]).

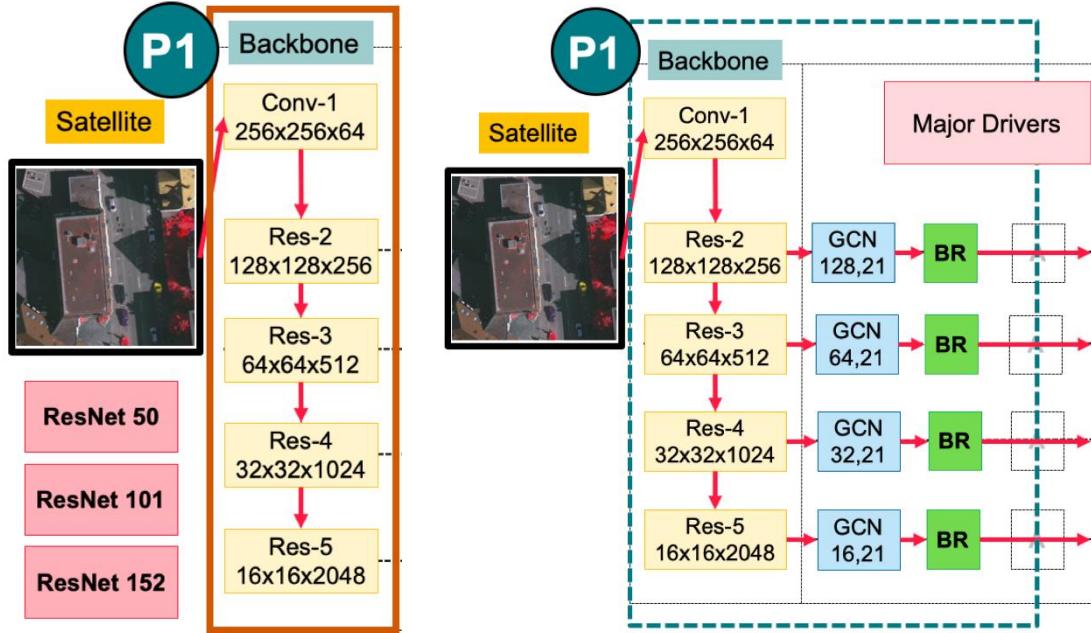


Figure 15. An overview of the whole backbone pipeline in (left) the main backbone with varying by ResNet50, ResNet101, and ResNet152; (right) the major drivers of our main classification network (composed of a global convolutional network (GCN) and a boundary refinement (BR) block [33])

4.3. The Channel Attention Block (A)

Attention mechanisms [14, 35, 36] in neural networks are very loosely based on the visual attention mechanism found in humans and equips a neural network with the ability to focus on a subset of its inputs (or features): it selects specific inputs. Human visual attention is well-studied, and while there are different models, all of them essentially come down to being able to focus on a certain region of an image with a very high resolution, perceiving the surrounding image in a medium resolution, and then adjusting the focal point over time.

To apply this attentional layer to our network, the channel attention block is shown in Block A in Figure 9 and its detailed architecture is shown in Figure 16. It is

designed to change the weights of the remote sensing features on each stage (level), so that the weights are assigned more values on important features adaptively.

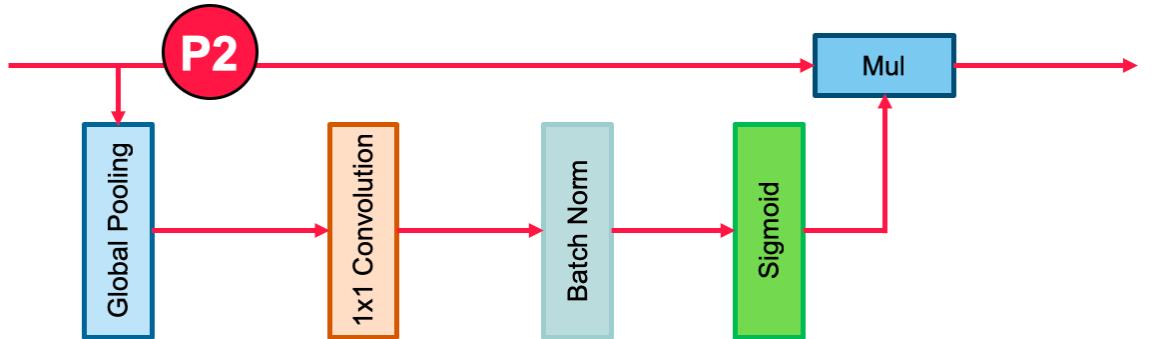


Figure 16. Components of the channel attention block. The red lines represent the down sample operators, respectively. The red line cannot change the size of feature maps. It is only a path for information passing

In the proposed architecture, a convolution operator gives the probability of each class at each pixel. In equation 7, the final score is summed over all channels of the feature maps.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

$$y_k = F(x; w) = \sum_{i=1, j=1}^D w_{i,j} x_{i,j} \quad (7)$$

Where x is the output feature of network. w represents the convolution's kernel, and $k \in \{1, 2, 3, 4, 5, 6, 7, \dots, K\}$. The number of channels is represented by K , and D is the set of pixel positions.

$$\delta_i(y_k) = \frac{\exp(y_k)}{\sum_{j=1}^k \exp(y_j)} \quad (8)$$

Where δ is the prediction probability. y_j is the output of the network. As shown in equations 7 and 8, the final predicted label is the category with the highest probability. Therefore, we suppose that the prediction result is y_0 of a certain patch, while its true label is y_1 . Therefore, we can introduce a parameter α to change the highest probability value from y_0 to y_1 , as equation 9 shows.

$$\bar{y} = \alpha y = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_k \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix} = \begin{bmatrix} \alpha_1 w_1 \\ \vdots \\ \alpha_k w_k \end{bmatrix} \times \begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix} \quad (9)$$

Where \bar{y} is the new prediction of the network, and $\alpha = \text{Sigmoid}(x; w)$.

Based on the above formulation of the Channel Attention Block, we can explore its practical significance. In equation 7, it implicitly indicates that the weights of different channels are equal. However, the features in different stages have different degrees of discrimination, which results in different consistency of prediction. Consequently, in equation 9, α value applies the feature maps x , which represents the feature selection with the channel attention block.

4.4. Domain-Specific Transfer Learning (TL)

The overall idea of transfer learning is to use knowledge learned from tasks for which many labeled data are usable in settings where only little-labeled data are available. Creating labeled data is expensive, so optimally leveraging an existing dataset is key. Certain low-level features, such as edges, shapes, corners, and intensity, can be shared across tasks, and new high-level features specific to the target problem can be learned [37-39]. Additionally, knowledge from an existing task acts as an additional input when learning a new target task.

Although the deep learning approach often performs promising prediction performance, it requires a large amount of training data. Since it is difficult to obtain annotated satellite images, the performance in prior works has been limited.

Fortunately, there is a recent concept called domain-specific transfer learning [37-39] that allows one to reuse the weights obtained from other domains' inputs. It is currently very popular in the field of deep learning because it enables one to train deep neural networks with comparatively insufficient data. This is very useful since most real-world problems typically do not have millions of labeled data points to train such complex models.

In terms of inadequacy, we propose an effective transfer deep neural network to perform knowledge transfer between a very high resolution (VHR) corpus and a medium resolution (MR) corpus. It is shown in Figure 17.

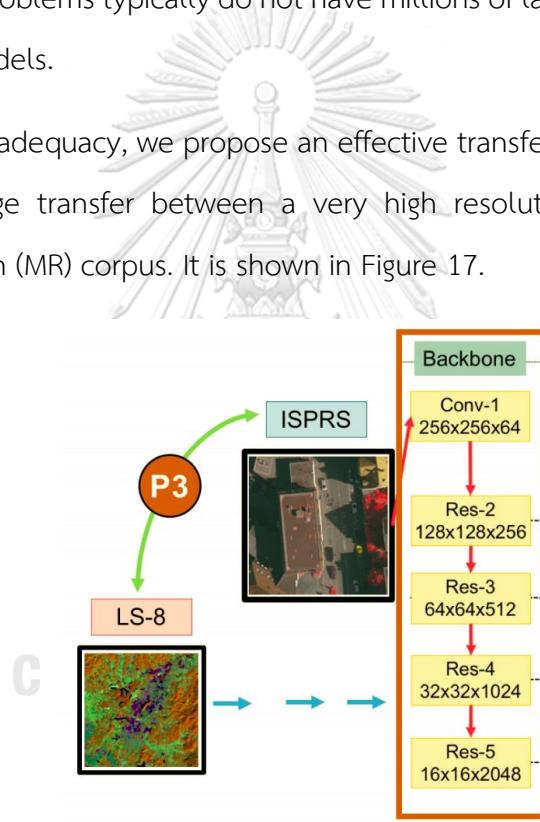


Figure 17. The domain-specific transfer learning strategy reuses pre-trained weights of models between two datasets—very high (ISPRS) and medium (Landsat-8; LS-8) resolution images

4.5. Feature Fusion Concept

The main idea of feature fusion is to fuse two paths are different in layer for captured by the backbone encodes mostly rich detail information and combine low and high features effectively on remote sensing tasks.

To apply this feature fusion concept [14, 69-73] to our network, the feature fusion block is shown in Block FF in Figure 14 and 15. Its detailed architecture is shown in Figure 18. It is designed to fuse the different in layer of feature representation for capture mostly rich detail information.

Inspired by the idea of feature fusion [14, 70-73] that integrates multiplication, additional, or concatenate layers. Convolution with 1×1 filters is used to transform features with different dimensions into the shape, which can be fused. The fusion method contains an addition process. Each layer of the backbone network such as VGG, Inception, ResNet, or HR creates the feature map for specific. We proposed to combine output with low-level features (front-end network) with the deep model and refine the feature information. As shown in Figure 18, the kernel maps after fusing will be calculated as Equation (10):

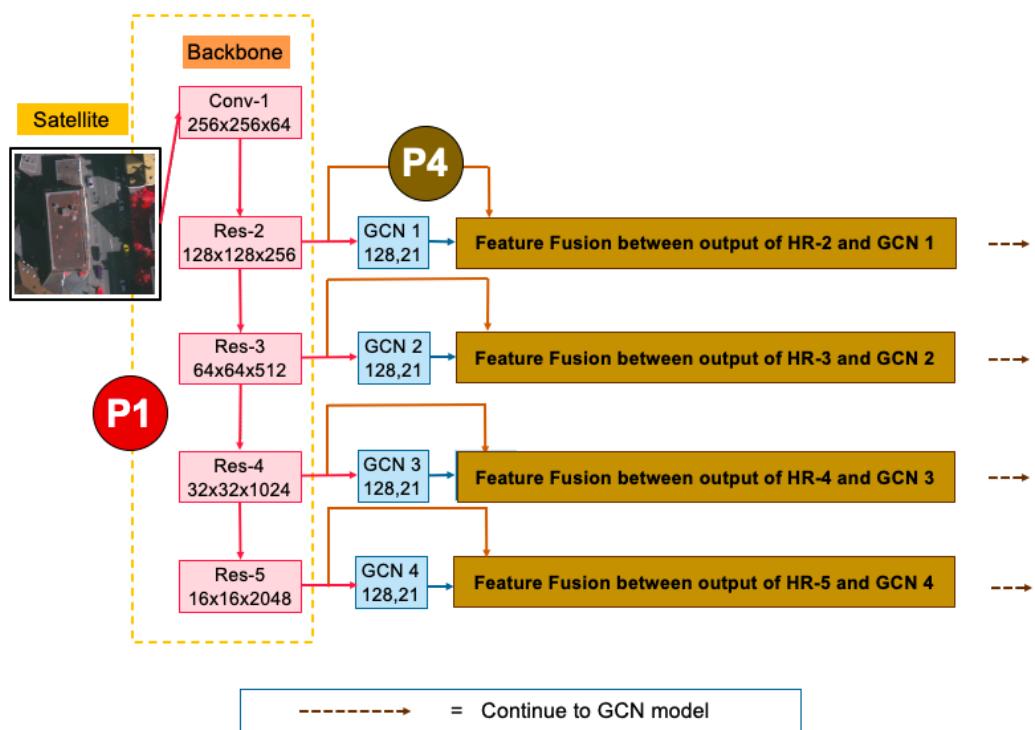


Figure 18. The framework of our feature fusion strategy.

$$Z_{add} = X_1 \oplus X_2 \oplus X_3 \dots \oplus X_i \dots \oplus X_j \quad (10)$$

Where j advert to the index of the layer, X_i is a set of output activation maps of one layer and \oplus advert to element-wise addition. Hence, the nature of the addition process encourages essential information to build classifiers to comprehend the feature details. It denotes all bands of Z_{add} to hold more feature information.

Hence, the nature of the addition process encourages essential information to build classifiers to comprehend the feature details. It denotes all bands of Z_{add} to hold more feature information.

Equation (11) shows the relationship between input and output. Thus, we take the fusion activation map into the model again, it can be performed as Equation (13):

$$\bar{y}^i = \text{ReLU}(w^T x^i + b) \quad (11)$$

Where x is the input and output of layer of the convolution recorded as \bar{y}^i ; b and w refer to bias and weight. The cost function in this work is demonstrated via Equation (12).

$$J(w, b) = -\frac{1}{m} \times [(1 - y^i) \log(1 - \bar{y}^i) + (y^i \log(\bar{y}^i))] \quad (12)$$

Where y refers to segmentation target of input (each image) and J , w , and b are the loss, weight, and bias value, respectively.

$$Y_{add} = f(W_k Z_{add} + B_k) \quad (13)$$

The feature fusion procedure always transforms into the same thing when using additional procedures. In this work, we use addition fusion elements, as shown in Figure 14.

4.6. Depthwise Atrous Convolution (DA)

Depthwise Atrous Convolution (DA) [11, 13, 17, 67, 75] is presented to settle the contradictory requirements between the larger region of the input space that affects a particular unit of the deep network (receptive fields) and activation map resolution.

DA is a robust operation to reduce the number of parameters (weights) in the layer of the CNN while maintaining a similar performance that includes the computation cost and tunes the kernel's field-of-view in order to capture a generalized standard convolution operation and multi-scale information. An atrous filter can be a dilated kernel in varied rates, e.g., rate = 1, 2, 4, 8, by inserting zeros into appropriate positions in the kernel mask.

Basically, the DA module uses atrous convolutions to aggregate multi-scale contextual information without dissipating resolution orderly in each layer. It generalizes “Kronecker-factored” convolutional kernels, and it allows for broad receptive fields, while only expanding the number of weights logarithmically. In other words, DA can apply the same kernel at distinct scales using various atrous factors.

Compared to the ordinary convolution operator, atrous (dilated) convolution is able to achieve a larger receptive field size without increasing the numbers of kernel parameters.

Our motivation is to apply DA to solve challenging scale variations and to trade off precision in aerial and satellite images, as shown in Figure 19.

In a one-dimensional (1D) case, let $x[i]$ denote input signal, and $y[i]$ denote output signal. The dilated convolution is formulated as Equation (14):

$$y[i] = \sum_{j=1}^J x[i + a \cdot k] \cdot w[j] \quad (14)$$

Where a is the atrous (dilated) rate, $w[j]$ denotes the j^{th} parameter of the kernel, and J is the filter size. This equation reduces to a standard convolution when $d = 1, 2, 4$, and 8 , respectively.

In the cascading mode from DeepLabV3 [11, 25] and Atrous Spatial Pyramid Pooling (ASPP) [17], multi-scale contextual information can be encoded by probing the incoming features with dilated convolution to capture sharper object boundaries by continuously recovering the spatial characteristic. DA has been applied to increase the computational ability and achieve the performance by factorizing a traditional convolution into a depth-wise convolution followed by a point-wise convolution, such as 1×1 convolution (it is often applied on the low-level attributes to decrease the whole of the bands (kernel maps)).

DA is found that it significantly decreases the computation complexity of the proposed model while maintaining comparable performance. There is 1×1 convolution on the low-level features before concatenation to reduce the number of channels since the corresponding low-level features usually contain a large number of channels, which may burden the importance of the rich encoder features.

To simplify notations, $H_{j,a}(x)$ is term of a dilated convolution, and ASPP can be performed as Equation (15).

$$y = H_{3,1}(x) + H_{3,2}(x) + H_{3,4}(x) + H_{3,8}(x) \quad (15)$$

To improve the semantics of shallow features, we apply the idea of multiple dilated convolution with different sampling rates to the input kernel map before continuing with the decoder network and adjusting the dilation rates (1, 2, 4, and 8) to configure the whole process of our proposed method called “GCN152-TL-A-FF-DA”, shown in P5 in Figures 14 and 19.

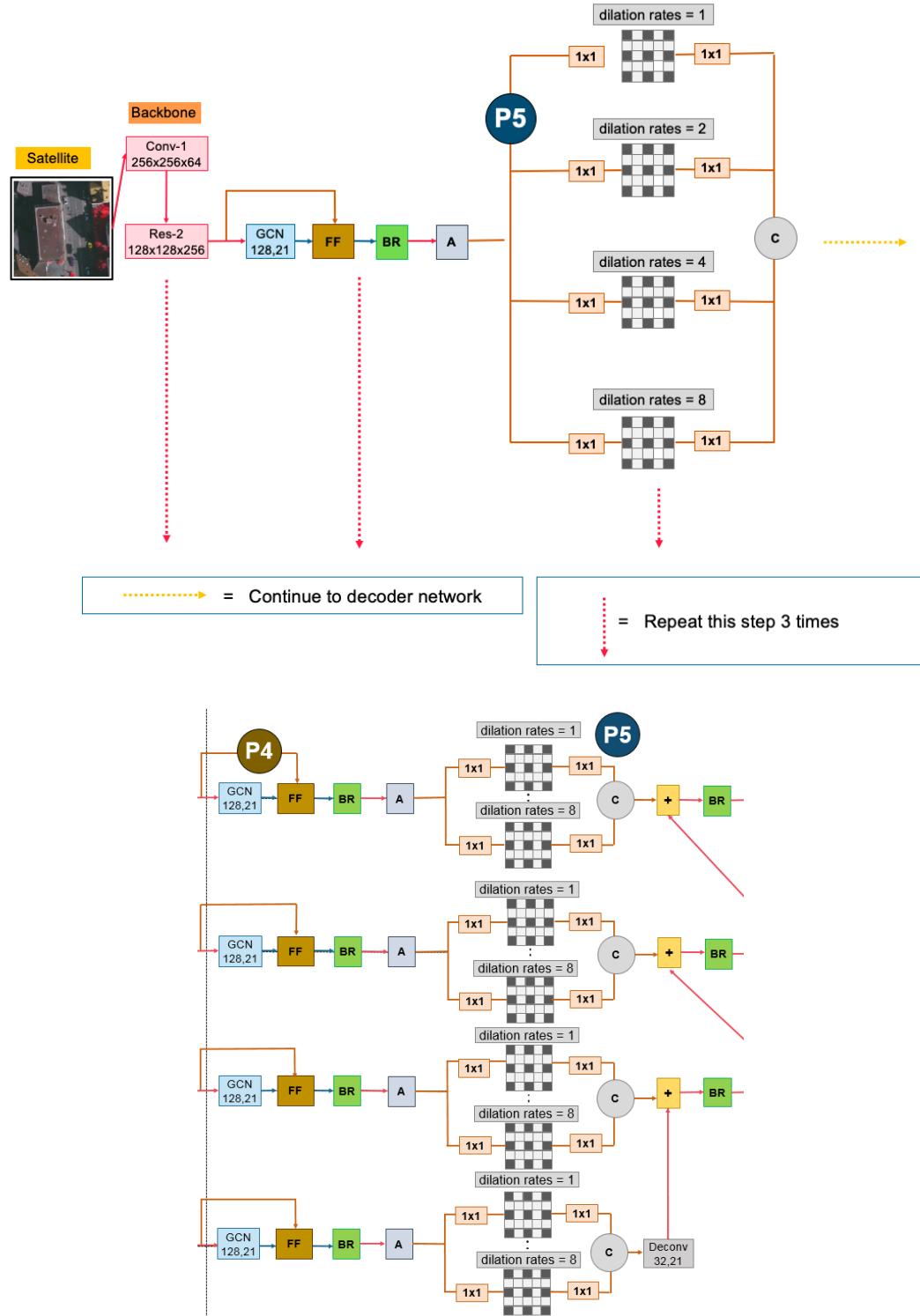


Figure 19. The Depthwise Atrous Convolution (DA) module in the proposed parallel pyramid method for improving feature fusion.

CHAPTER V

EXPERIMENTAL RESULTS

In our experiments, there are two main sources of data: public and private corpora. The private corpora is the medium resolution imagery received from the satellite “Landsat-8” used by the government organization in Thailand named GISTDA (Geo-Informatics and Space Technology Development Agency (Public Organization)). Since there are two variations of annotations, the Landsat-8 data is considered as two data sets: one with three classes and the other with five classes, as shown in Table 4. The public corpora is very high-resolution imagery from the standard benchmark called “ISPRS Vaihingen (Stuttgart)”. Evaluations based on classification/segmentation metrics, e.g., F1 Score, Precision, Recall and Average Accuracy are deployed with all experiments.

Table 4 Abbreviations on our Landsat-8 corpora.

Abbreviation	Description
Landsat-8w3c corpus	Landsat-8 corpus with 3 classes
Landsat-8w5c corpus	Landsat-8 corpus with 5 classes

5.1. Landsat-8w3c Dataset

Landsat-8 is an American earth observation satellite and it collects and archive medium resolution (30-m spatial resolution) multispectral image data affording seasonal coverage of the global landmasses for a period of no less than 5 years. Landsat-8 [29] images consist of nine spectral bands with a spatial resolution of 30 m for Bands 1–7 and 9. The ultra-blue Band 1 is useful for coastal and aerosol studies. Band 9 is useful for cirrus cloud detection. The resolution for Band 8 (panchromatic) is 15 m. Thermal Bands 10 and 11 are useful in providing more accurate surface

temperatures and are collected at 100 m. The approximate scene size is 170 km north-south by 183 km east-west (106 mi by 114 mi). Since Landsat-8 data includes additional bands, the combinations used to create RGB composites differ from Landsat 7 and Landsat 5. For instance, Bands 4, 3, and 2 are used to create a color infrared (CIR) image using Landsat 7 or Landsat 5. To create a CIR composite using Landsat 8 data, Bands 5, 4, and 3 are used.

For this corpus, all images are taken in the area of the Northeast of Thailand (Isan), consists of 20 provinces in the northeastern region of Thailand and this region boasts high biodiversity and many endemic species, with several national parks. They are Thailand's largest region. The data set is made from the Landsat-8 satellite consisting of 1,420 satellite images, some samples are shown in Figure 20. This data set contains a massive collection of medium resolution imagery of (20,921 x 17,472) pixels. There are three classes: Para rubber (red), pineapple (green), and corn (yellow). From a total of 1,390 images, the images are separated into 1,000 training and 230 validation images, as well as 190 test images to compare with other baseline methods.

5.2. Landsat-8w5c Dataset

This data set is the same corpus from Landsat-8. Still, all images are taken in the area of the Nan province, is one of Thailand's seventy-seven provinces (Changwat) lies in upper northern Thailand, and to the north, it borders Sainyabuli of Laos. This province is in the remote Nan River valley, surrounded by forested mountains. It is annotated with five class labels: agriculture (yellow), forest (green), miscellaneous (brown), urban (red), and water (blue), as shown in Figure 21. There are 1,012 medium resolution satellite images of 17,200 x 16,300 pixels. From the total 1,039 images, the images are separated into 700 training and 239 validation images, as well as 100 test images to comparison to other baseline methods.

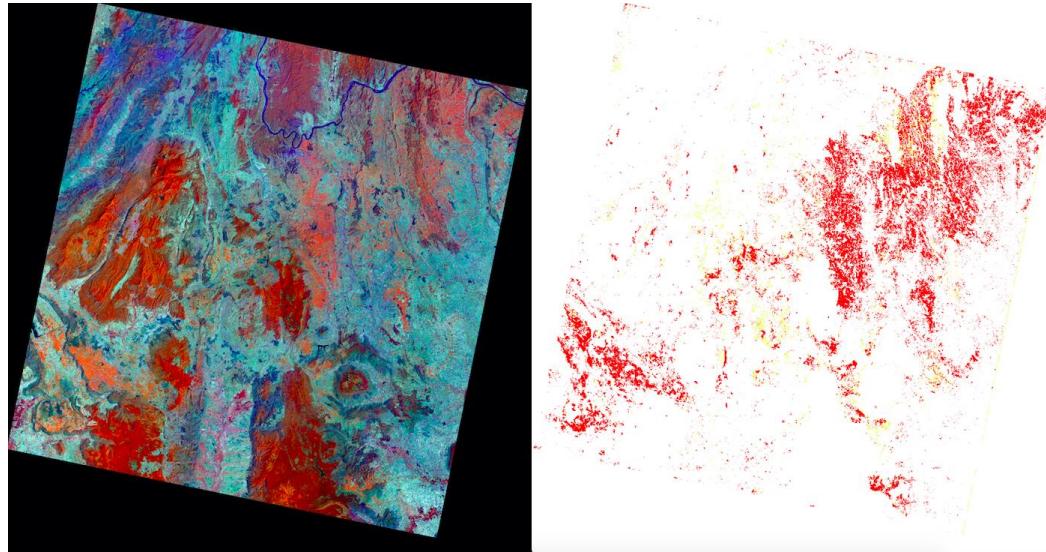


Figure 20. Sample satellite images from Isan (Northeastern Thailand), a zone in Thailand (**left**), and corresponding ground truth (**right**). The label of this data set includes three categories: corn (yellow), pineapple (green), and rubber tree (red)

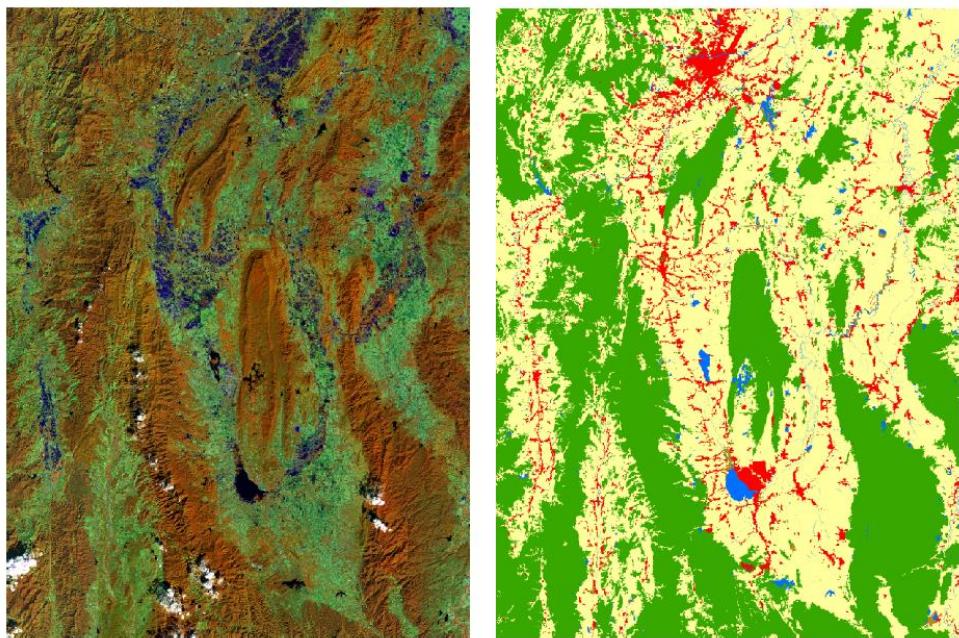


Figure 21. Sample satellite images from Nan, a province in Thailand (**left**), and corresponding ground truth (**right**). The label of medium resolution dataset includes five categories: agriculture (yellow), forest (green), miscellaneous (brown), urban (red), and water (blue)

5.3. ISPRS Vaihingen Dataset

One of the major challenges in remote sensing is the automated extraction of urban objects from data acquired by airborne sensors. The Semantic Labeling Contest provides two state-of-the-art airborne image corpora. The Vaihingen corpus shows a relatively small village with many detached buildings and small multi-story buildings, and the Potsdam corpus shows a typical historic city with large building blocks, narrow streets, and dense settlement structure. In our experiments, the Vaihingen corpus was selected and used.

The ISPRS 2D Semantic labeling challenge in Vaihingen [1, 2, 29] (Figure 22 and 23) was used as our benchmark dataset. It consists of three spectral bands (i.e., red, green, and near-infrared bands), the corresponding DSM (digital surface model) and the NDSM (normalized digital surface model) data. Overall, there are 33 images of about $2,500 \times 2,000$ pixels at a ground sampling distance (GSD) of about 9 cm in the image data. Among them, the ground truth of only 16 images are available, and those of the remaining 17 images are withheld by the challenge organizer for the online test. For offline validation, we randomly split the 16 images with ground truth available into a training set of 10 images and a validation set of 6 images. For this work, DSM and NDSM data in all experiments on this dataset were not used. Following other methods, four tiles (Image Numbers 5, 7, 23, and 30) were removed from the training set as the validation set. Experimental results are reported on the validation set if not specified.

This challenge of ISPRS semantic segmentation at Vaihingen (Stuttgart) is used to be our standard corpus. They were captured over Vaihingen in Germany. The data set is a subset of the data used for the test of digital aerial cameras carried out by the German Association of Photogrammetry and Remote Sensing (DGPF).



Figure 22. Overview of the ISPRS 2D Vaihingen Labeling corpus. There are 33 tiles.

Numbers in the figure refer to the individual tile flag

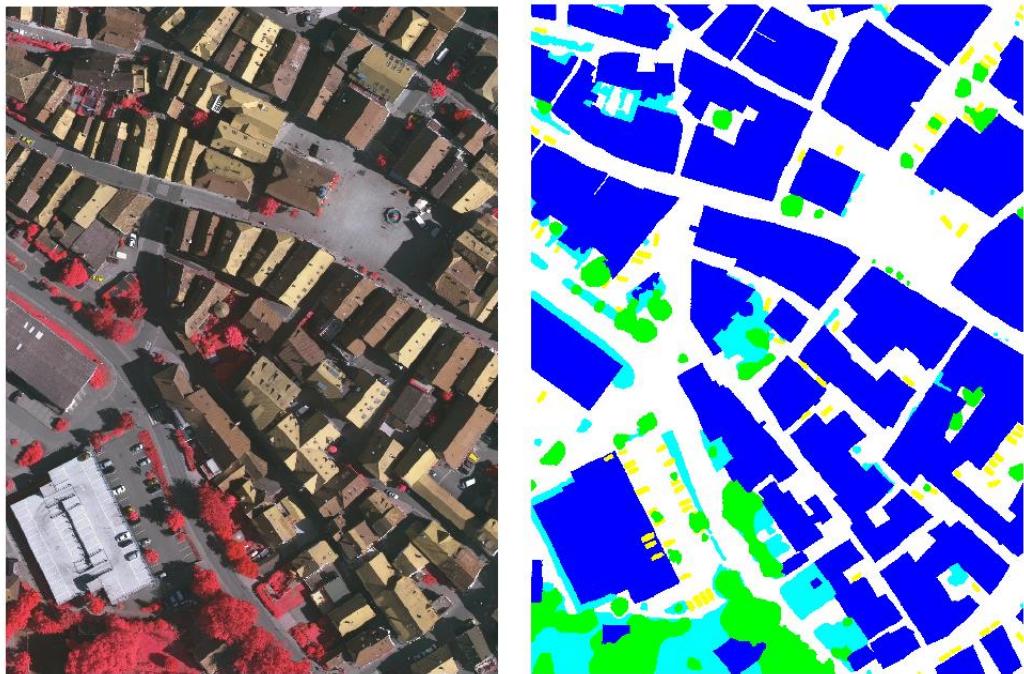


Figure 23. The sample input tile from Figure 7 (**left**) and corresponding ground truth (**right**). The label of the Vaihingen Challenge includes six categories: impervious surface (imp surf, white), building (blue), low vegetation (low veg, cyan), tree (green), car (yellow), and clutter/background (red)

5.3. Evaluation

The multi-class classification task can be considered as multi-segmentation, where class pixels are positives and the remaining non-spotlight pixels are negatives. Let TP denote the number of true positives, TN denote the number of true negatives, FP denote the number of false positives, and FN denote the number of false negatives.

Precision, recall, and F1-score are shown in equations 10-13. Precision is the percentage of correctly classified main pixels among all predicted pixels by the classifier. Recall is the percentage of correctly classified main pixels among all actual main pixels. F1 is a combination of precision and recall.

To evaluate the performance of different deep models, we will discuss the above one major metrics (F1) on each category, and the mean value of metrics to assess the average performance.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (10)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (11)$$

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (12)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (13)$$

5.4. Experimental Setup

The implementation is based on a deep learning framework, called "Tensorflow-Slim" [76], which is extended from Tensorflow. All experiments were conducted on servers with an Intel® Xeon® Processor E5-2660 v3 (25M Cache, 2.60 GHz), 32 GB of

memory (RAM), an Nvidia GeForce GTX 1070 (8 GB), an Nvidia GeForce GTX 1080 (8 GB), and an Nvidia GeForce GTX 1080 Ti (11 GB). Instead of using the whole image (1500×1500 pixels) to train the network, we randomly cropped all images to be 512×512 as inputs of each epoch.

For training, the RMSPropOptimizer optimizer [60] was chosen with an initial learning rate of 0.004 and the weight decay of 0.00001. Batch normalization [59] is used before each convolutional layer in our implementation to ease the training and make it be able to concatenate feature maps from different layers. To avoid overfitting, common data augmentations are used as details in Chapter 4.1. For measurements, we use the F1-score as the metric.

Inspired by [14, 19, 35, 36, 48], we use the "poly" learning rate policy where the learning rate is multiplied by equation (14) with a power of 0.9 and an initial learning rate as 4×10^{-3} . The learning rate is scheduled by multiplying the initial as seen in equation (14).

$$\text{learning rate} = \text{init learning rate} * \left(1 - \frac{\text{epoch}}{\text{MaxEpoch}}\right)^{\text{power}} \quad (14)$$

All models are trained for 30 and 50 epochs with a mini-batch size of 4, and each batch contains the cropped images that are randomly selected from training patches. These patches are resized to 521×521 pixels. The statistics of BN is updated on the whole mini-batch.

This chapter illustrates the details of our experiments. The proposed deep learning network is based on the GCN with three improvements: (i) varying the backbones using ResNet, (ii) channel attention and global average pooling, (iii) domain-specific transfer learning, (iv) feature fusion, and (v) depthwise atrous convolution. From all proposed strategies, there are eight acronyms of strategies as shown in Table 5.

Table 5. Abbreviations on our proposed deep learning methods

Abbreviation	Description
A	Channel Attention Block
GCN	Global Convolutional Network
GCN50	Global Convolutional Network with ResNet50
GCN101	Global Convolutional Network with ResNet101
GCN152	Global Convolutional Network with ResNet52
TL	Domain-Specific Transfer Learning
FF	Feature Fusion Module
DA	Depthwise Atrous Convolution

For the experimental setup, there were three experiments on two remotely sensed datasets: the Landsat-8 dataset and the ISPRS Vaihingen Challenge dataset (details in Chapter 5.1, Chapter 5.2, and Chapter 5.3). The experiments aimed to illustrate that each proposed strategy can improve performance. First, the GCN152 method was compared to the GCN50 method and the GCN101 method for the diverse backbones using ResNet with different numbers of layers on the GCN network strategy. Second, the GCN152-A method was compared to the GCN152 method for the channel attention strategy. Third, the pre-full proposed technique GCN152-TL-A method was compared to existing methods for the concept of domain-specific transfer learning. Next, the proposed technique GCN152-TL-A-FF method for fusing each level feature from the backbone model and the global model of GCN to enrich local and global features. And last, the full proposed technique GCN152-TL-A-FF-DA proposes to bridge the semantic gap and implement durable multi-level feature aggregation to extract complementary information from very shallow features. In the end, we will call our full proposed method as "**Encoders Matter**" for the concept to fuse low-level features and high-level features to boost the performance of a rare class of remote sensing corpus.

5.5. Results of the Landsat-8w5c Corpus with Discussion

In this subsection, the Landsat-8w5c corpus was conducted on all experiments. We compare “**GCN152-TL-A-FF-DA**” [74] network (column (f)) to CNN baselines via Tables 6 and 7. “**GCN152-TL-A-FF-DA**” is the winner with a F1 of 0.9361. Furthermore, it is also the winner in all classes especially water and urban class that are composed with low-level features. More detailed results are described in the next subsection and are presented in Tables 6 and 7 for the results of this data set, Landsat-8w5c. It is shown that our network with all strategies, GCN152-TL-A-FF (Encoders Matter), outperforms other methods. More details will be discussed to show that each of the proposed techniques can improve accuracy. Only in this experiment is there a state-of-the-art baseline, including a deep convolutional encoder-decoder (DCED) [1, 12, 30, 31].

Table 6. Results of the testing data of the Landsat-8 (**Nan**) corpus between baseline and 7 variations of proposed techniques in terms of precision, recall, and F1-score

Method	Pretrained	Backbone	Model	Pre	Re	F1
Baseline	-	-	DCED [1]	0.857	0.894	0.874
Proposed	-	Res50	GCN	0.881	0.872	0.875
	-	Res101	GCN	0.862	0.897	0.877
	-	Res152	GCN	0.892	0.878	0.884
	-	Res152	GCN-A	0.907	0.929	0.917
	ISPRS	Res152	GCN-A	0.921	0.918	0.918
	ISPRS	Res152	GCN-A-FF	0.930	0.924	0.927
	ISPRS	Res152	GCN-A-FF-DA	0.934	0.939	0.936

5.5.1. The Effect of an Enhanced GCN on the Landsat-8w5c Corpus

For Nan corpus, our first strategy aims to increase an F1-score of the network by varying backbones using ResNet 50, ResNet 101, and ResNet 152 rather than the traditional one, the DCED method. From Table 7, the F1 of GCN152 (0.884) outperforms that of GCN50 (0.877), GCN101 (0.875), and the baseline method, DCED (0.874); this yields a higher F1 at 1%, 0.3%, and 0.1%, respectively.

Table 7. Results of the testing data of Landsat-8 corpus (**Nan**) between each class with our proposed techniques in terms of accuracy

Method	Model	Agri	Forest	Misc	Urban	Water
Baseline	DCED [1]	0.982	0.962	0.763	0.854	0.725
Proposed	GCN50	0.967	0.948	0.817	0.881	0.792
	GCN101	0.976	0.929	0.685	0.929	0.785
	GCN152	0.976	0.950	0.823	0.913	0.797
	GCN152-A	0.984	0.944	0.882	0.899	0.822
	GCN152-TL-A	0.974	0.953	0.864	0.934	0.828
	GCN152-TL-A-FF	0.986	0.982	0.918	0.956	0.844
	GCN152-TL-A-FF-DA	0.989	0.957	0.934	0.949	0.868

The main reason is due to higher precision, but a slightly lower recall. This can imply that enhanced GCN is more significantly efficient than the DCED method (baseline) for this medium resolution corpus and ResNet with a large number of layers is more robust than the small number of layers.

When comparing the results between the original GCN method and the enhanced GCN methods on the Landsat-8 corpus (Table 6 to Table 7), it is clearly shown that a

GCN with a larger layer of backbone can improve network performance in terms of F1-score.

5.5.2. The Effect of Using Channel Attention on the Landsat-8w5c Corpus

For Nan corpus, our second mechanism focused on applying the channel attention block (details in Chapter 4.3 to change the weights of the features on each stage to enhance consistency. In Table 6, the F1 of GCN152-A (0.917) is greater than that of GCN152 (0.884); this yields a higher F1-score at 3.3%.

The result (Figure 24d) shows that can make the network to obtain discriminative features stage-wise to make the prediction intra-class consistent. This is based on the consideration that we re-weighted all feature maps of each layer.

5.5.3. The Effect of Using Domain-Specific Transfer Learning on LS8w5c Corpus

For Nan corpus, this strategy aims to use approach of domain-specific transfer learning (details in Chapter 4.4) by reusing the pre-trained weight from the GCN152-A model on the ISPRS Vaihingen corpus. From Table 6 and Table 7, the F1 of the GCN152-TL-A method is the winner if compare with previous proposed methods; it clearly outperforms not only the baseline but also all previous generations.

Its F1 is higher than that of the DCED (baseline) at 4.4% on Nan corpus. Additionally, the result illustrates that the concept of domain-specific transfer learning can enhance recall (2.4%) of Nan corpus.

5.5.4. The Effect of Using Feature Fusion on Landsat-8w5c Corpus

Our strategy aims to use approach of domain-specific transfer learning (details in Chapter 4.5) by fusing two paths that different in layer of feature representation. From Table 6, the F1 of the GCN152-TL-A-FF method is the winner; it clearly outperforms not only the baseline but also all previous generations. Its F1 is higher than that of the

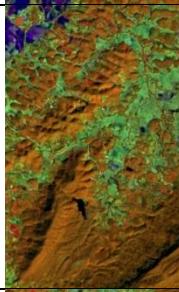
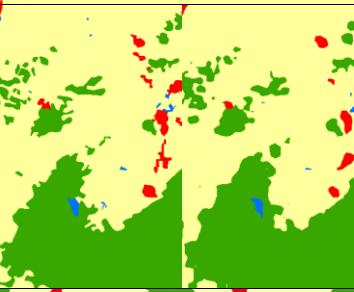
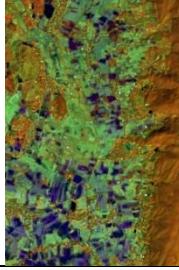
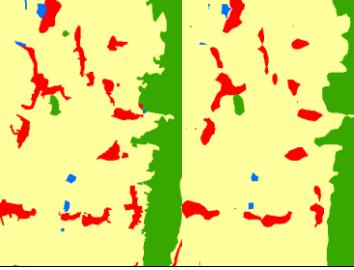
DCED (baseline) at 5.1%. Additionally, the result illustrates that the concept of domain-specific transfer learning can enhance both precision (7.3%) and recall (3.0%) on Nan corpus.

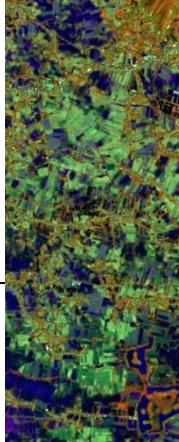
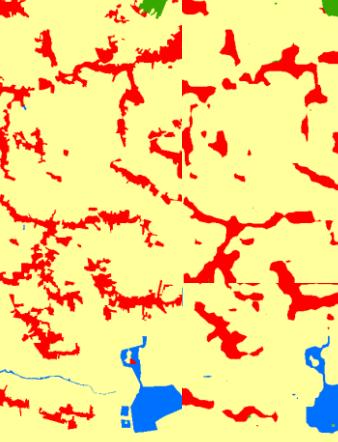
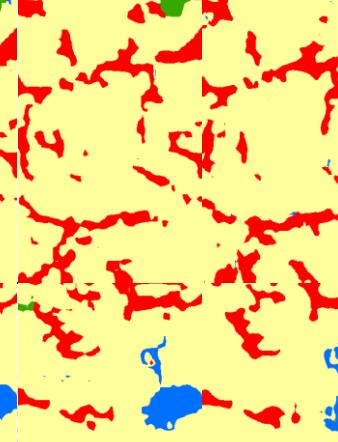
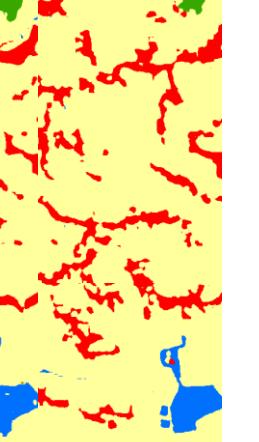
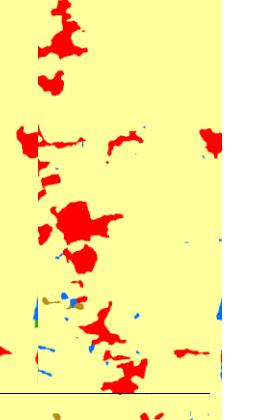
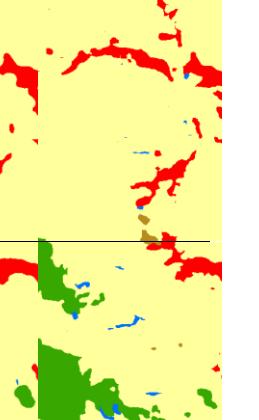
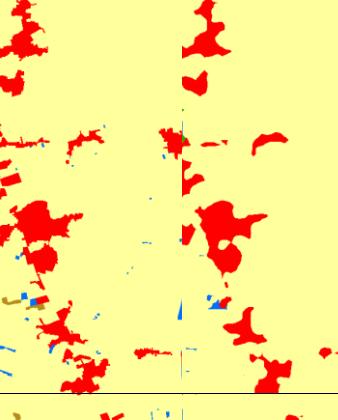
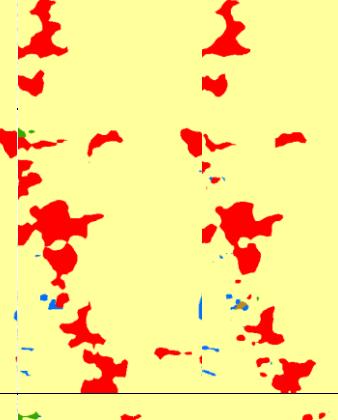
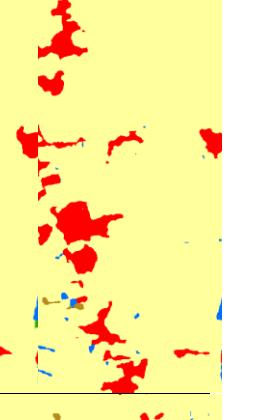
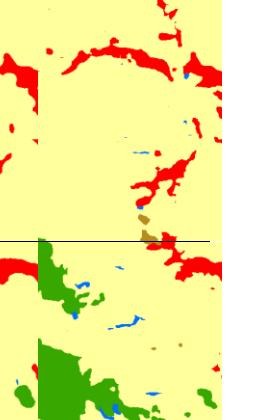
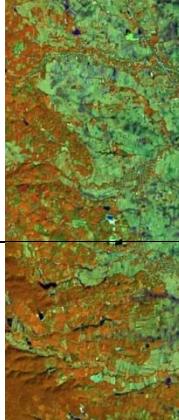
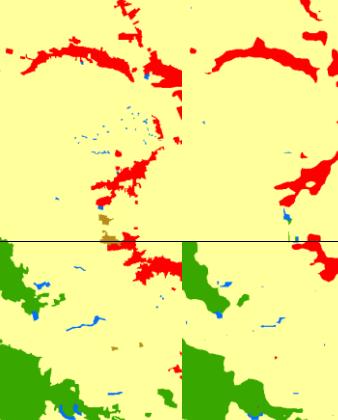
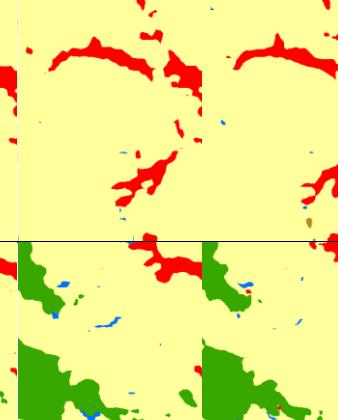
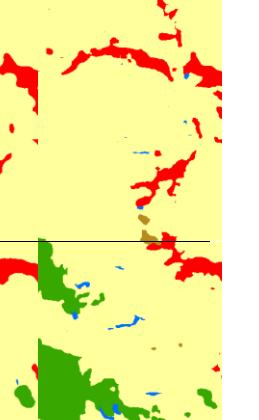
5.5.5. The Effect of Using Depthwise Atrous Convolution on Landsat-8w5c Corpus

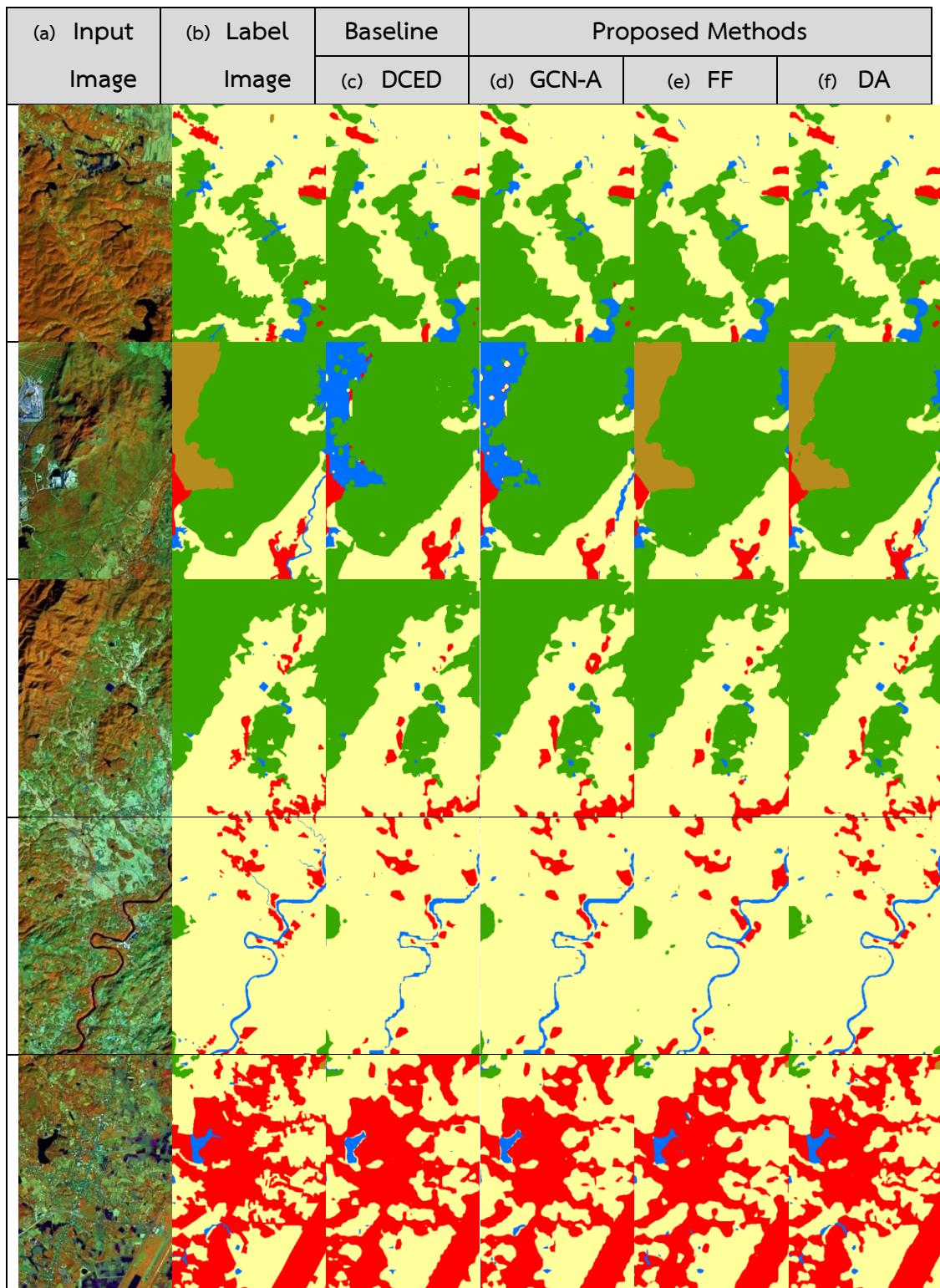
The last policy points to the performance of the “Depthwise Atrous Convolution” method by enhancing the features of CNN for improving the previous step (details in Chapter 4.6). The F1 score of the “GCN152-TL-A-FF-DA” approach is the conqueror. It is more eminent than DCED and GCN152-TL-FF at 6.2% and 0.9%, consecutively, shown in Tables 6 and 7. In the dilated convolution, filters are boarder, which can capture better overview details resulting in (i) larger coverage areas and (ii) connected small areas together. For an analysis of each class, our final model is clearly the winner in all classes with accuracy beyond 95% in two classes: agriculture and urban classes. Figures 24 show ten sample outputs from our proposed methods (column (d to f)) compared to the baseline (column (c)) to expose improvements in its results, and that founds that Figures 24f is likewise to the ground images. Our investigation found that the dilated convolutional concept can make our model have better overview information so that it can capture larger areas of data.

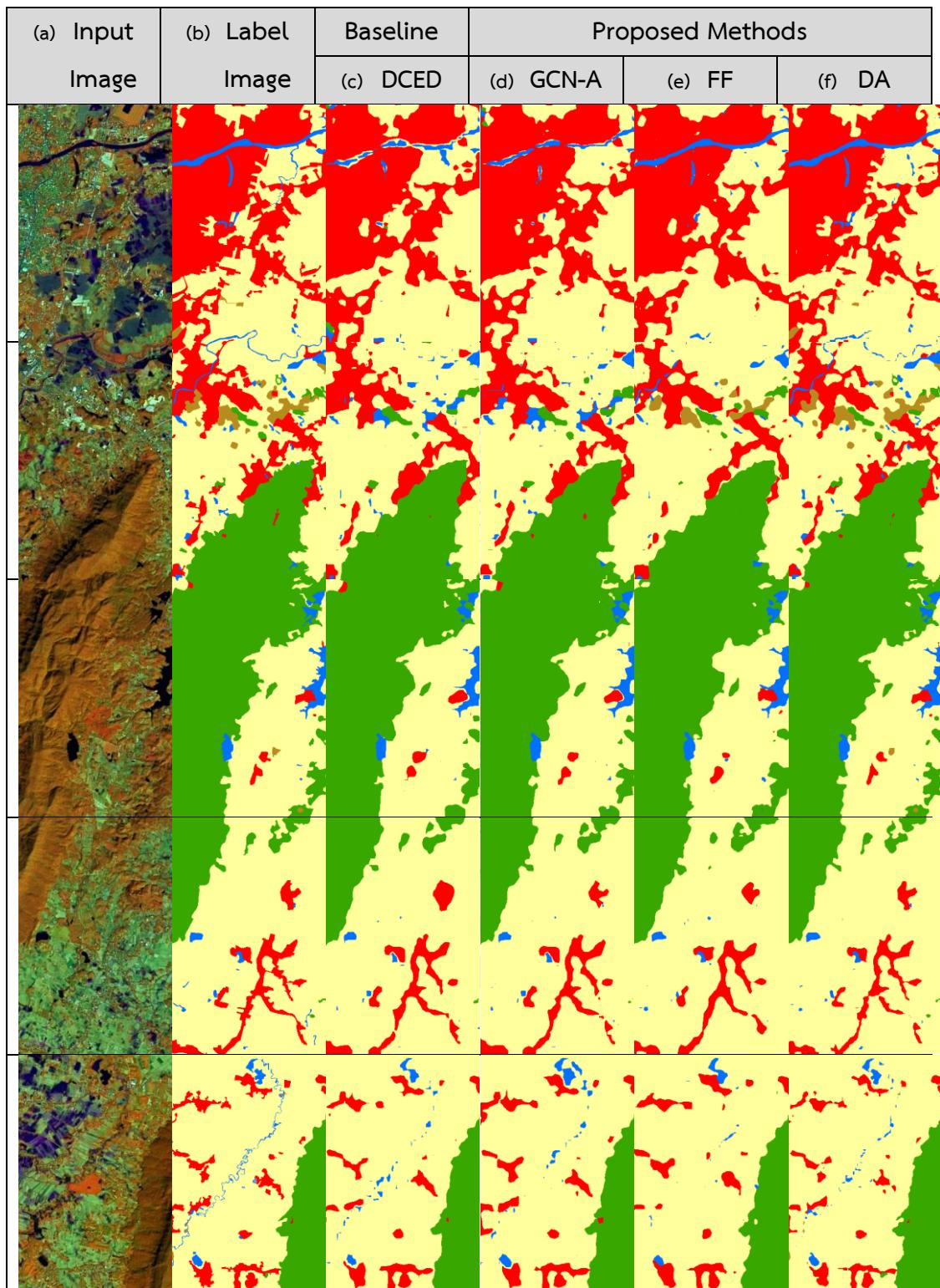
Figure 24 show ten sample results from the proposed method. By applying all strategies, the images in the last column (Figure 24f) are similar to the ground truths (Figure 24b). Furthermore, F1- is improved for each strategy we added to the network as shown in Figure 24c-f and Figure 24c-f.

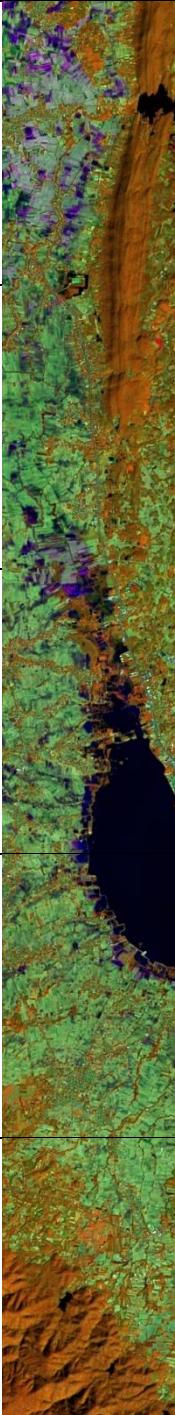
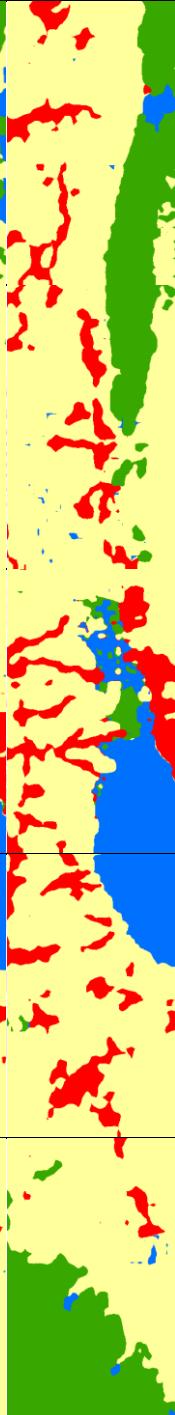
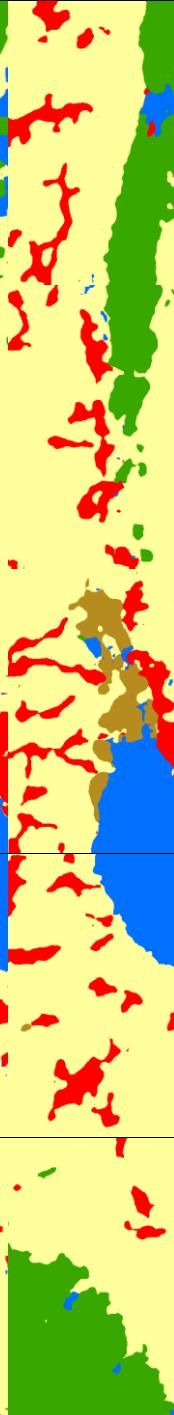
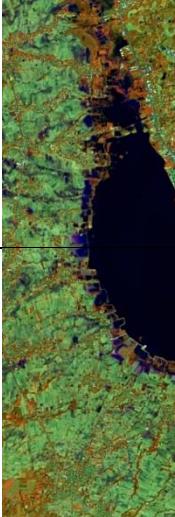
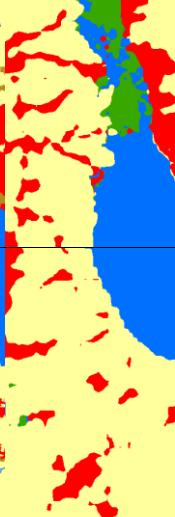
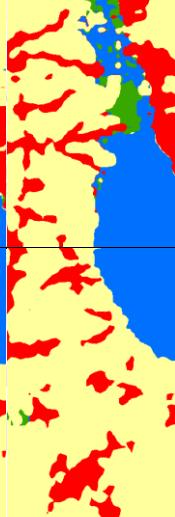
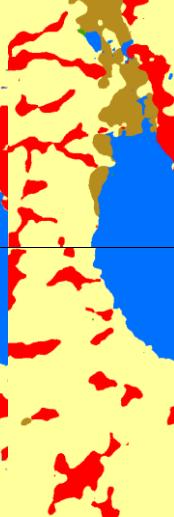
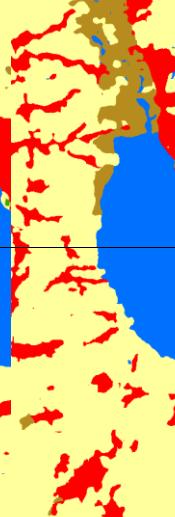
From our investigation, we found that since the dilated convolutional concept can make our model have better overview information so that it can capture larger areas of data.

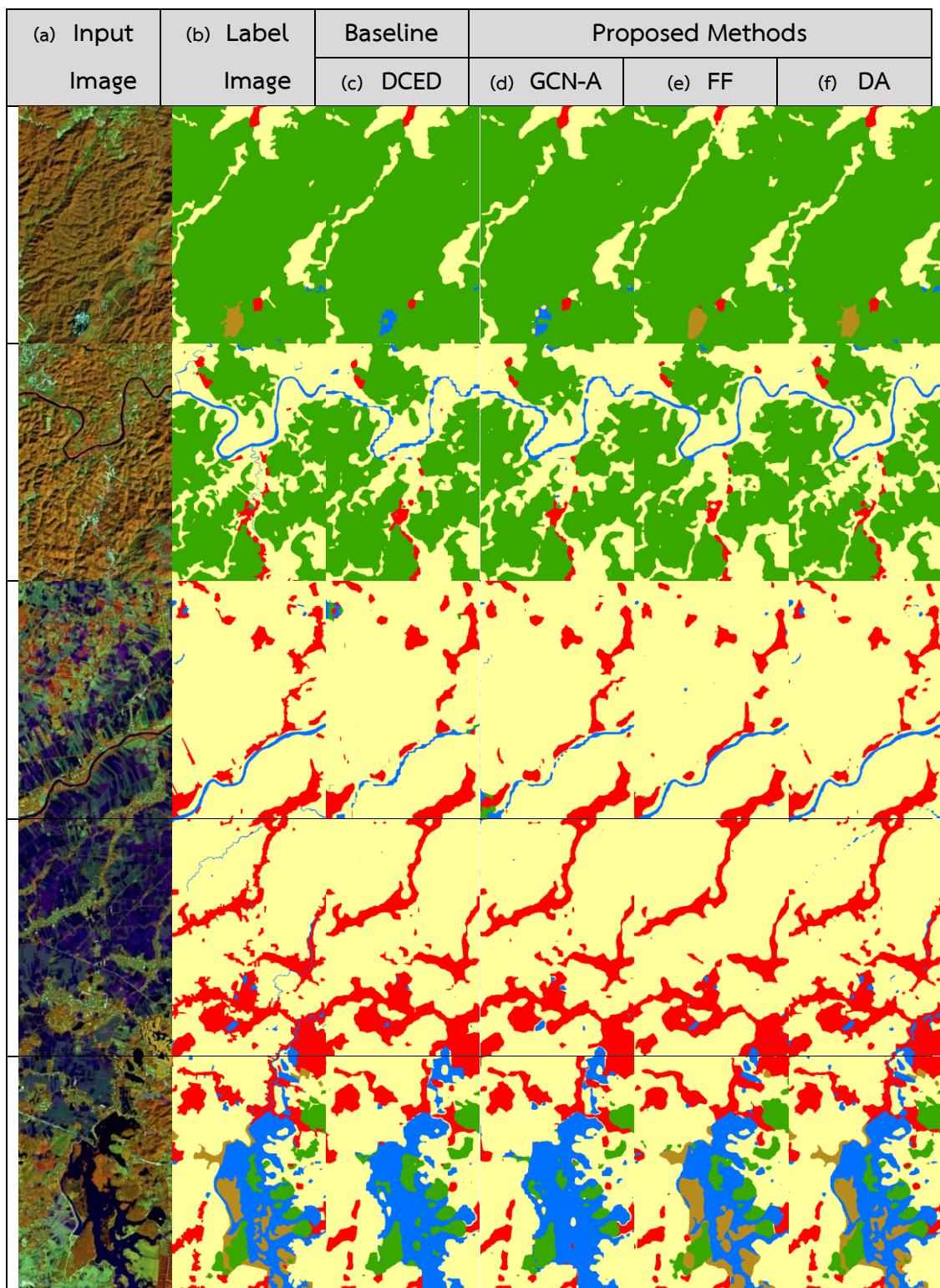
(a) Input Image	(b) Label Image	Baseline	Proposed Methods			
			(c) DCED	(d) GCN-A	(e) FF	(f) DA
						
						
						

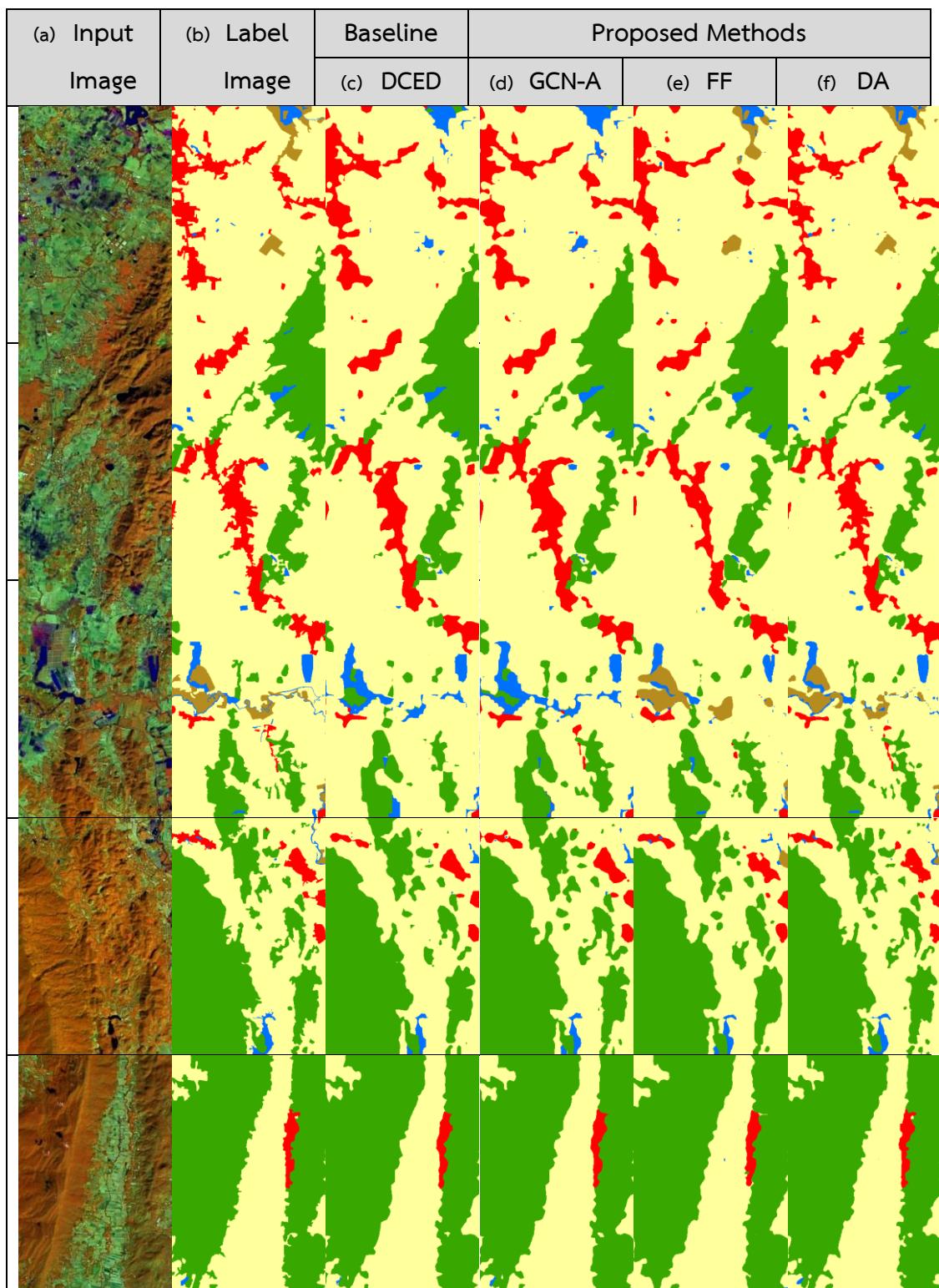
(a) Input Image	(b) Label Image	Baseline	Proposed Methods			
			(c) DCED	(d) GCN-A	(e) FF	(f) DA
						
						
						





(a) Input Image	(b) Label Image	Baseline	Proposed Methods			
			(c) DCED	(d) GCN-A	(e) FF	(f) DA
						
						
						





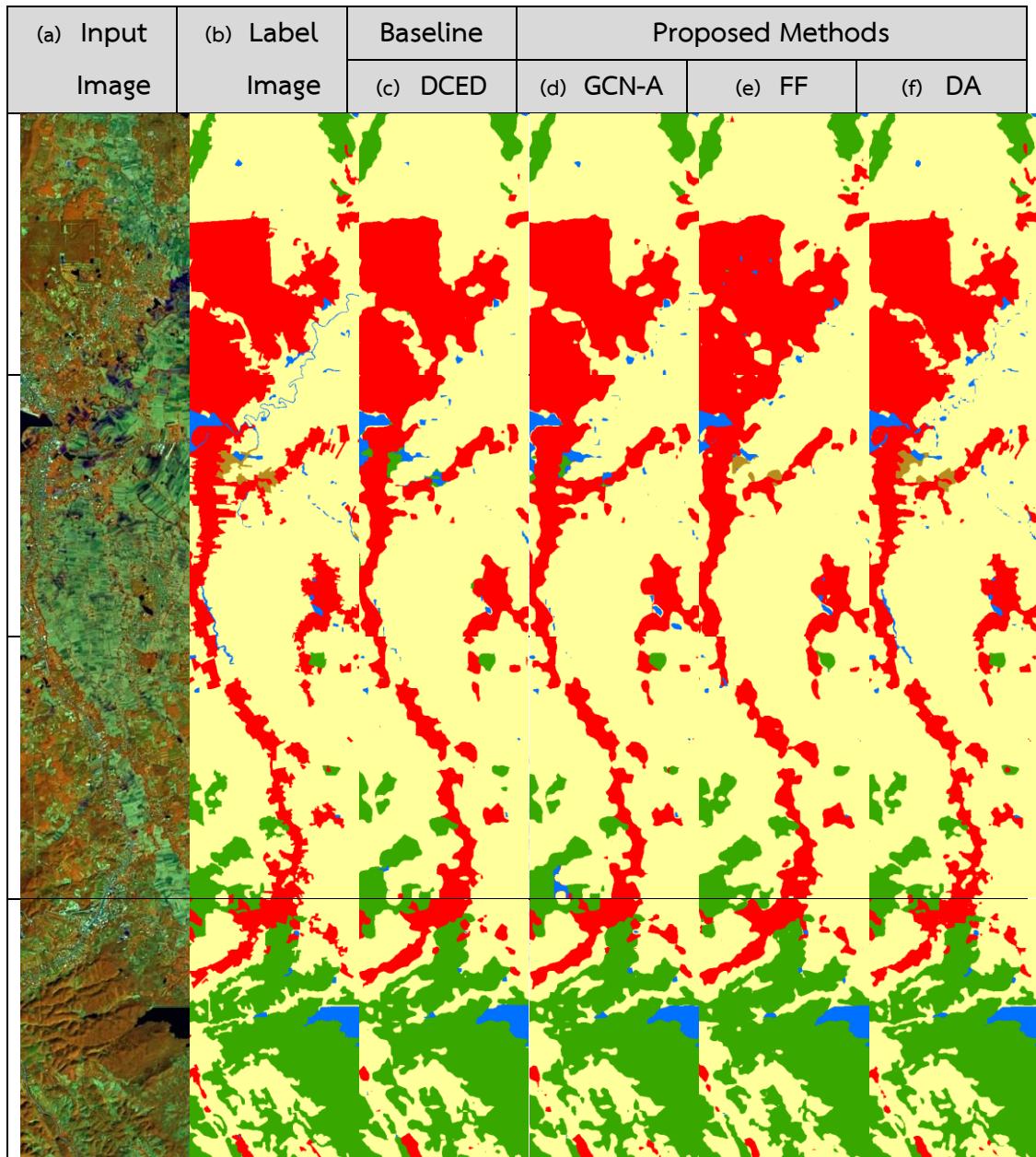


Figure 24. Testing sample inputs and output satellite images on Landsat-8w5c in the Nan province in Thailand, where rows refer to different images. (a) Original input image. (b) Target map (ground truth). (c) Output of Encoder–Decoder (Baseline). (d) Output of GCN152-TL-A. (e) Output of GCN152-TL-A-FF. and (f) Output of GCN152-TL-A-FF-DA. The label of medium resolution dataset includes five categories: Agriculture (yellow), Forest (green), Miscellaneous (Misc, brown), Urban (red) and Water (blue)

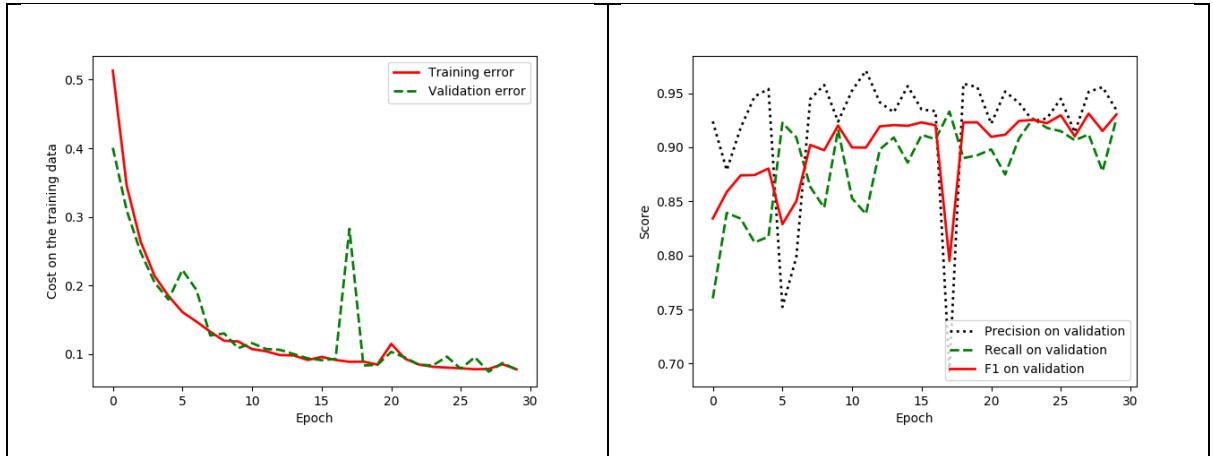


Figure 25. Graph (learning curves) on Landsat-8w5c data set of the proposed approach, “GCN152-TL-A-FF-DA”; x refers to epochs, and y refers to different measures (a) Plot of model loss (cross-entropy) on training and validation corpora; (b) performance plot on the validation corpus.

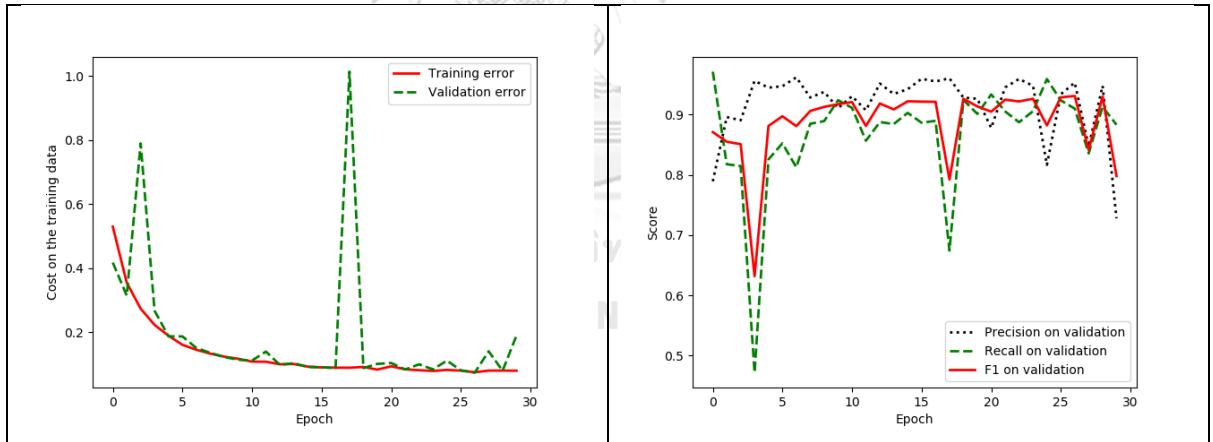


Figure 26. Graph (learning curves) on Landsat-8w5c data set of the baseline approach, DCED [1, 12, 30, 31]; x refers to epochs, and y refers to different measures (a) Plot of model loss (cross-entropy) on training and validation corpora; (b) performance plot on the validation corpus.

Ten sample testing results (shown as Figure 24) are based on the proposed method with respect to Nan (one of the Northern provinces (changwat) of Thailand and where agriculture is the main industry. The results of the last column look closest to the ground truth in the second column.

As can be seen in Figure 24, the performance of our best model outperforms other advanced models by a considerable margin on each category, especially for the agriculture (agri), miscellaneous (misc), and water classes. Furthermore, the loss and F1 curves shown in Figure 25a-b exhibit that our model performs better on all given categories (Figure 26a-b).

5.6. Results of the Landsat-8w3c Corpus with Discussion

The Landsat-8w3c corpus was used in all experiments. We distinguished between the alterations of the proposed approaches and CNN baselines. “GCN152-TL-A-FF-DA”, the full proposed method, is the winner with F1 of 0.9114. Furthermore, it is also the winner of all classes. More detailed results are given in the next subsection. Presented in Tables 8 and 9 are the results of this corpus, Landsat-8w3c.

5.6.1. The Effect of an Enhanced GCN on the Landsat-8w3c Corpus

For Isan corpus, our first strategy also aims to increase an F1-score of the network by varying backbones using ResNet 50, ResNet 101, and ResNet 152 rather than the traditional one, the DCED method. From Table 8, the F1 of GCN152 (0.876) outperforms that of GCN50 (0.874), GCN101 (0.872), and the baseline method, DCED (0.810); this yields a higher F1 at 6.6%, 6.4%, and 6.2%, respectively.

5.6.2. The Effect of Using Channel Attention on the Landsat-8w3c Corpus

For Isan corpus, our second mechanism also focused on applying the channel attention block (details in Chapter 4.3 to change the weights of the features on each stage to enhance consistency. In Table 8, the F1 of GCN152-A (0.877) is greater than that of GCN152 (0.876); this yields a higher F1-score at 0.1%.

5.6.3. The Effect of Using Domain-Specific Transfer Learning on LS8w3c Corpus

For Isan corpus, this strategy also aims to use approach of domain-specific transfer learning (details in Chapter 4.4) by reusing the pre-trained weight from the GCN152-A model on the ISPRS Vaihingen corpus. From Table 8 and Table 9, the F1 of the GCN152-TL-A method is also the winner if compare with previous proposed methods; it also clearly outperforms not only the baseline but also all previous generations.

Its F1 is higher than that of the DCED (baseline) model at 6.9% on Isan corpus. Additionally, the result illustrates that the concept of domain-specific transfer learning can enhance precision (1.7%) of Isan corpus.

For the analysis of each class, GCN152-TL-A achieved an average accuracy on Para rubber, pineapple, and corn for 0.869, 0.789, and 0.909, consecutively. Compared to DCED, it won in two classes: Para rubber and corn. However, it won against our previous work (GCN152-TL) on every class.

5.6.4. The Effect of Using Feature Fusion on Landsat-8w3c Corpus

From Table 9 and Table 10, the F1 of the GCN152-TL-A-FF method is the winner; it clearly outperforms not only the baseline but also all previous generations. Its F1 is higher than that of the DCED (baseline) at 8.9%. Additionally, the result illustrates that the concept of domain-specific transfer learning can enhance both precision (2.8%) and recall (4.5%) on Isan corpus.

It is interesting that the FF module can really improve the performance in all classes, especially in the Para rubber and pineapple classes. It outperforms both DCED and all previous baselines in all classes. To further investigate the results, Figures 27e shows that the model with FF can capture pineapple (green area) surrounded in Para rubber (red area).

5.6.5. The Effect of Using Depthwise Atrous Convolution on Landsat-8w3c Corpus

The last strategy aims to use an approach of “Depthwise Atrous Convolution” (details in Section 4.6) by extracting complementary information from very shallow features and enhancing the deep features for improving feature fusion of the Landsat-8w3c corpus. The “GCN152-TL-A-FF-DA” method is the victor. F1 is obviously more distinguished than DCED at 10.1% and GCN152-TL-A-FF at 1.3%, as shown in Tables 8 and 9.

For an analysis of each class, our model is clearly the winner in all classes with accuracy beyond 90% accuracy in two classes: Para rubber and corn. Figures 27 shows nine sample outputs from our proposed methods (column (d to f)) compared to the baseline (column (c)) to expose improvements in its results. From our investigation, we found that the dilated convolutional concept can make our model have better overview information so that it can capture larger areas of data.

There is a lower discrepancy (peak) in the validation data of “GCN152-TL-A-FF-DA”, Figure 28a) than in the baseline, Figure 29a. Moreover, Figures 29b and 28b show three learning graphs: precision, recall, and F1 lines. The loss graph of the “GCN152-TL-A-FF-DA” model seems flattered (very smooth) than the baseline in Figure 29a. The epoch at number 27 was selected to be a pre-trained model for testing and transfer learning procedures.

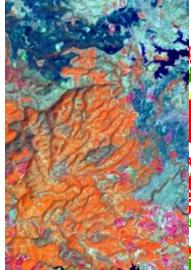
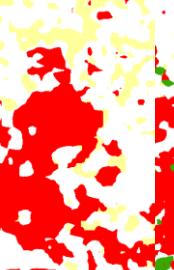
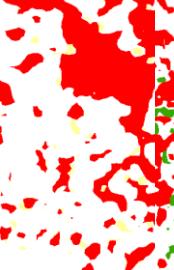
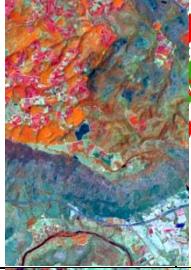
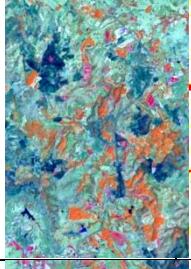
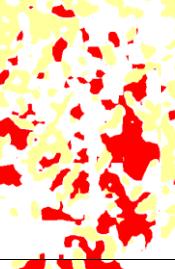
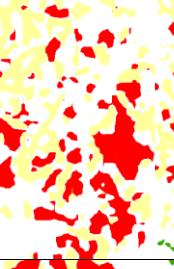
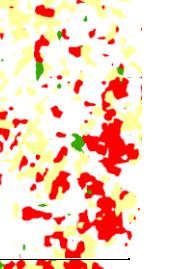
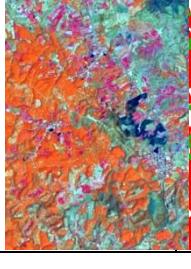
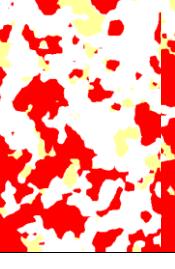
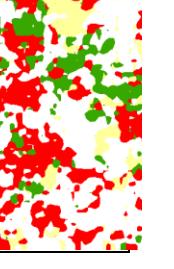
In the Landsat-8w3c corpus, for an analysis of each class, our model is clearly the winner in all classes with accuracy beyond 90% in two classes: Para rubber and corn. From our investigation, we found that since the dilated convolutional concept.

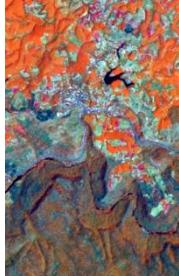
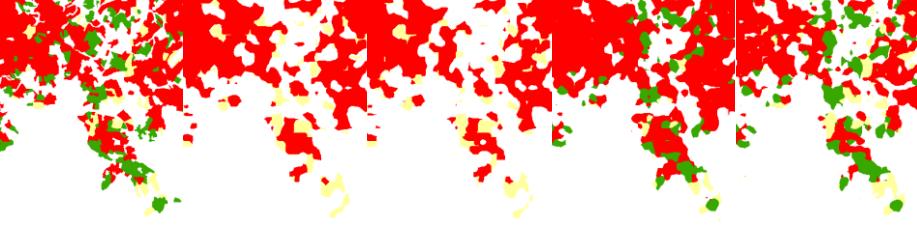
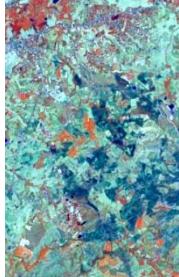
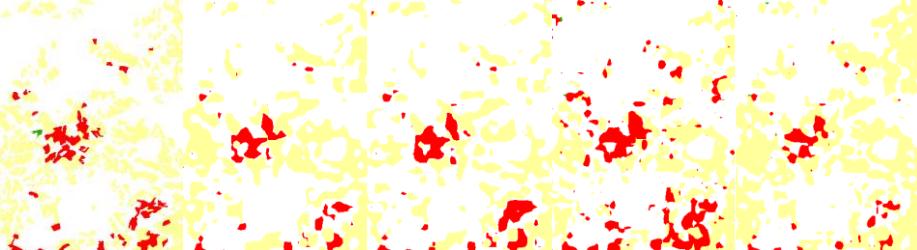
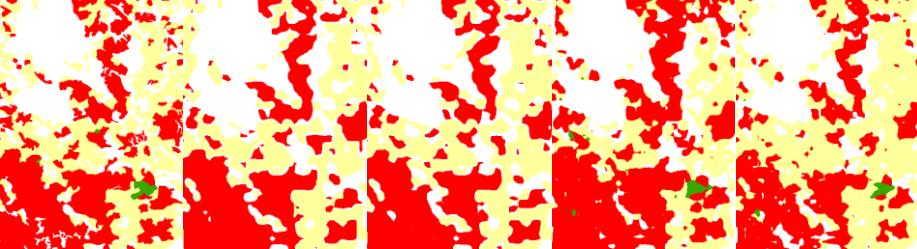
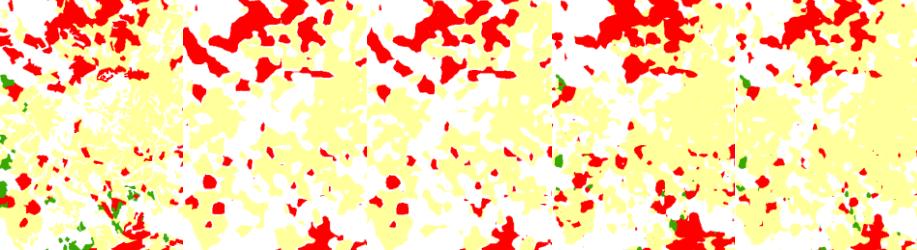
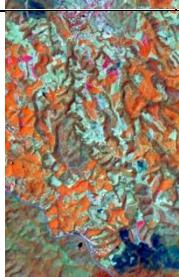
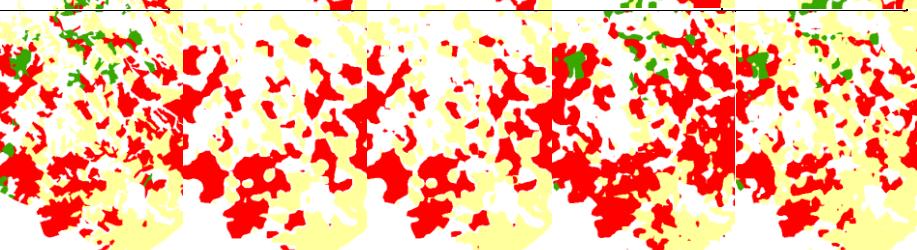
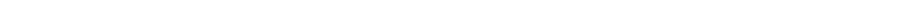
Table 8. Results of the testing data of the Isan corpus between baseline and 6 variations of proposed techniques in terms of precision, recall, and F1-score

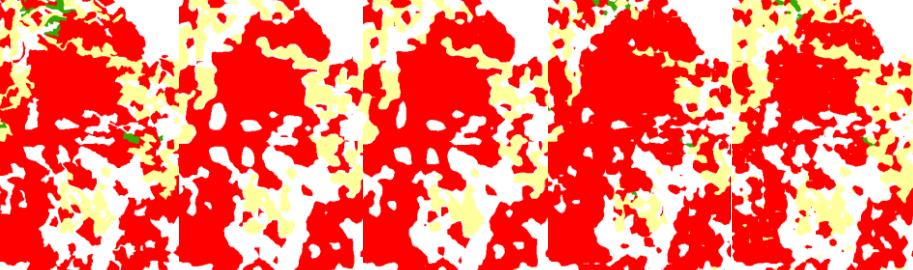
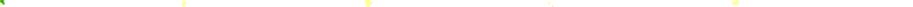
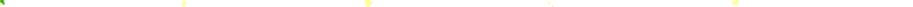
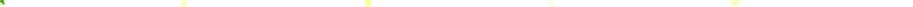
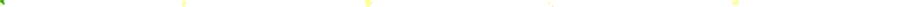
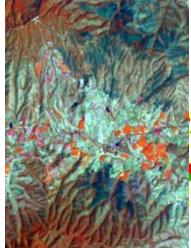
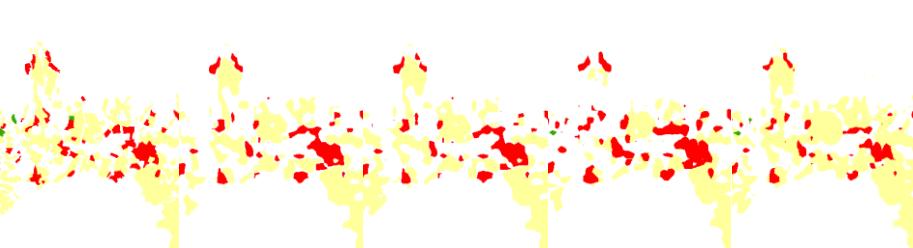
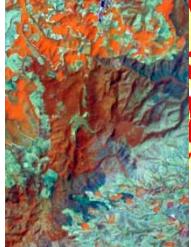
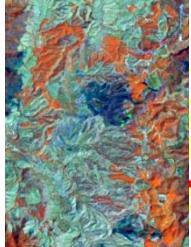
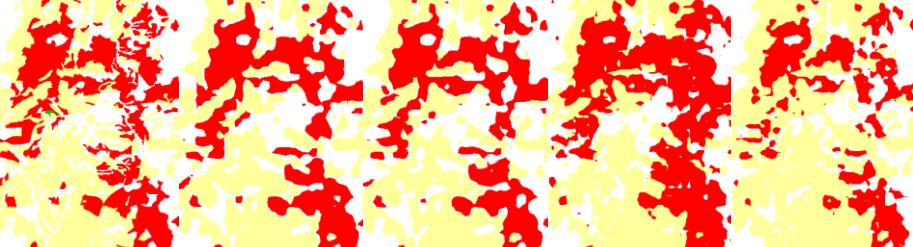
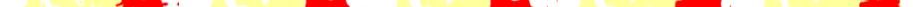
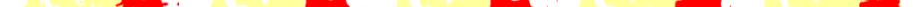
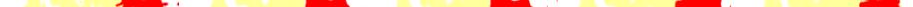
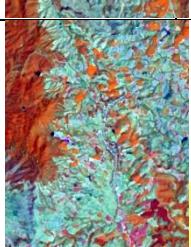
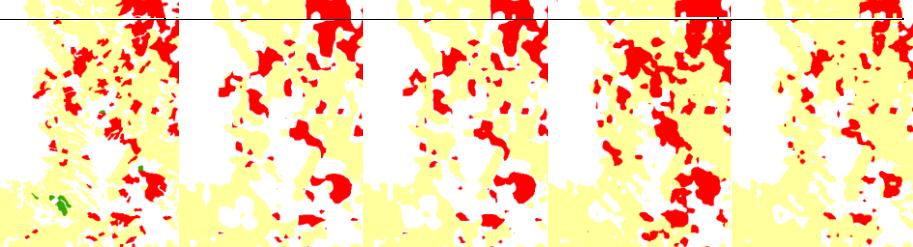
Method	Pretrained	Backbone	Model	Pre	Re	F1
Baseline	-	-	DCED [1]	0.861	0.782	0.810
Proposed	-	Res50	GCN	0.873	0.872	0.872
	-	Res152	GCN	0.860	0.898	0.876
	-	Res152	GCN-A	0.865	0.891	0.877
	ISPRS	Res152	GCN-A	0.878	0.883	0.879
	ISPRS	Res152	GCN-A-FF	0.889	0.914	0.899
	ISPRS	Res152	GCN-A-FF-DA	0.900	0.923	0.911

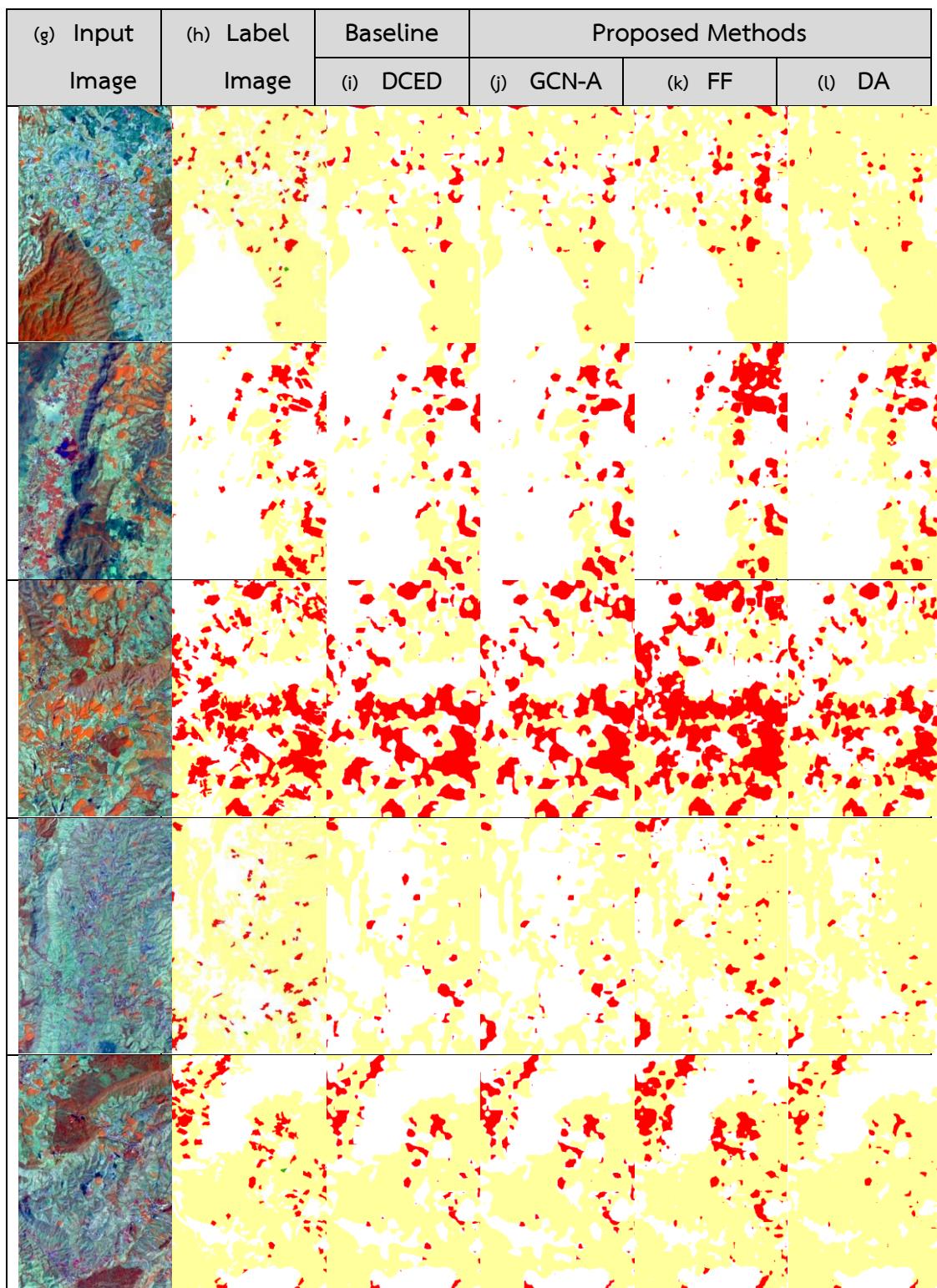
Table 9. Results of the testing data of Landsat-8 (**Isan**) corpus between each class with our proposed techniques in terms of accuracy

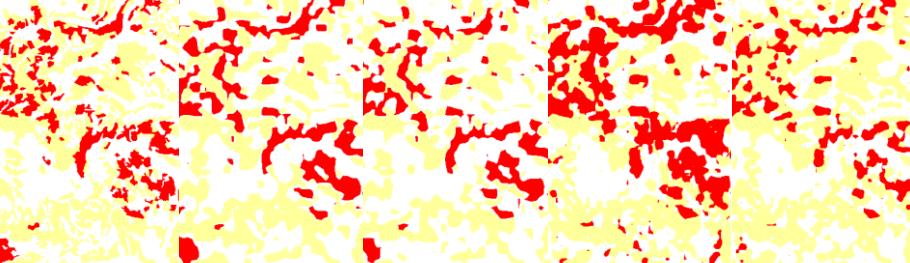
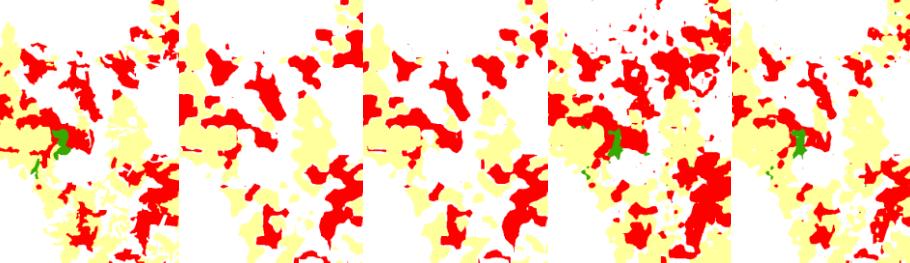
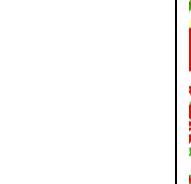
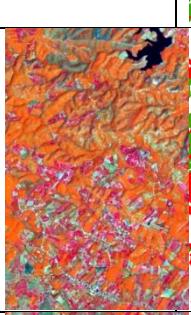
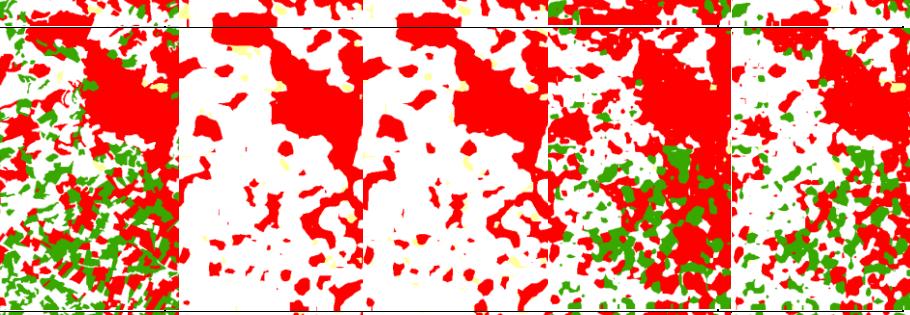
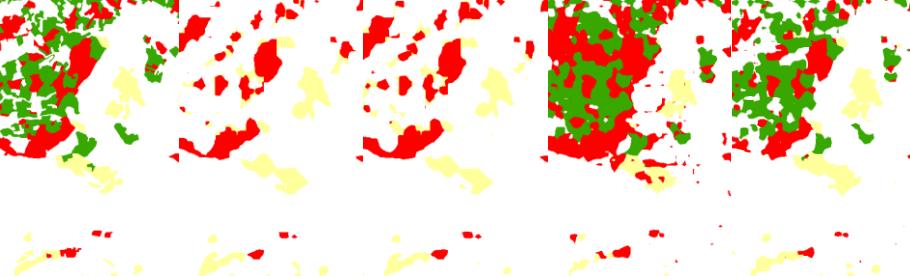
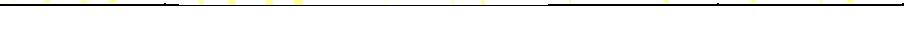
Method	Model	Corn	Pineapple	Rubber tree
Baseline	DCED [1]	0.861	0.782	0.810
Proposed	GCN50	0.873	0.872	0.872
	GCN101	0.865	0.884	0.874
	GCN152-A	0.865	0.891	0.877
	GCN152-TL-A	0.878	0.883	0.879
	GCN152-TL-A-FF	0.889	0.914	0.899
	GCN152-TL-A-FF-DA	0.949	0.868	0.898

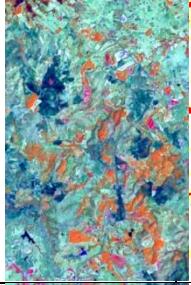
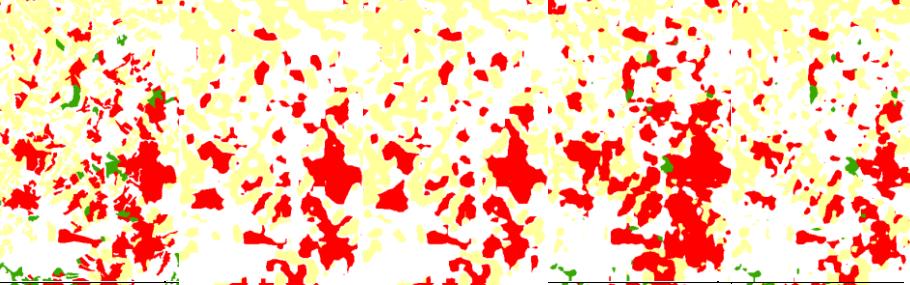
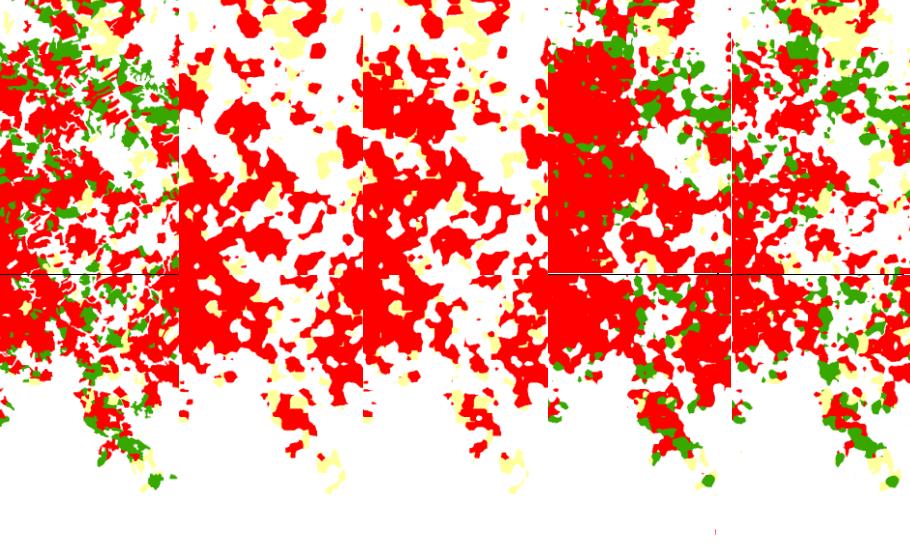
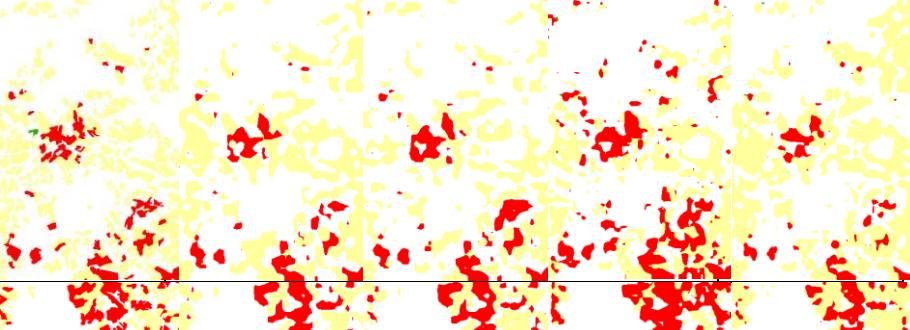
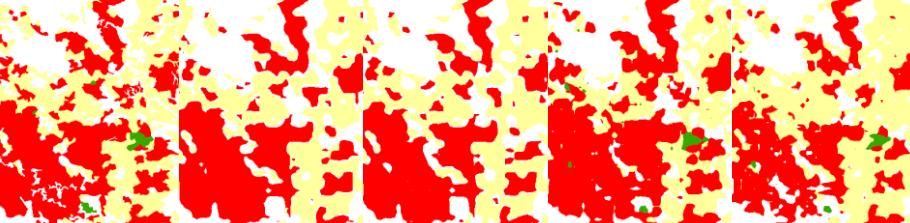
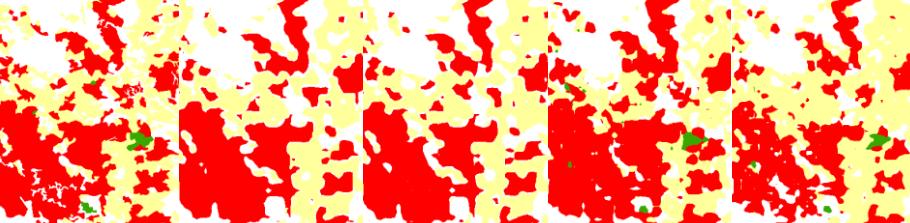
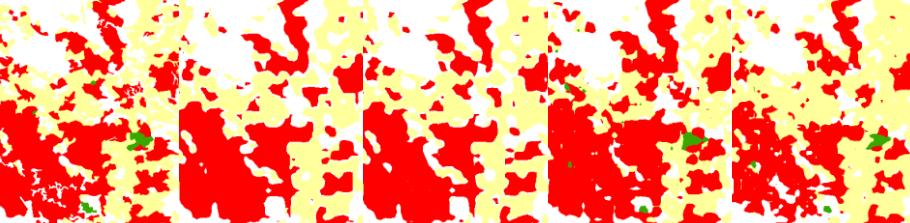
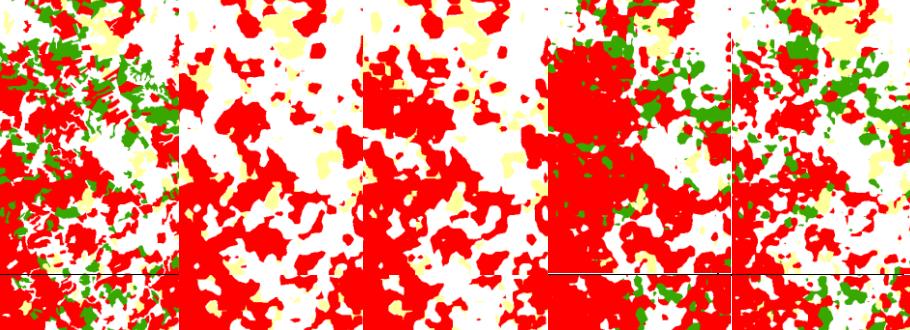
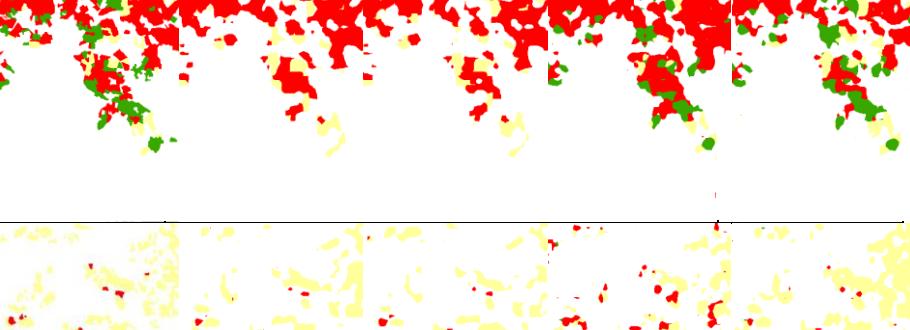
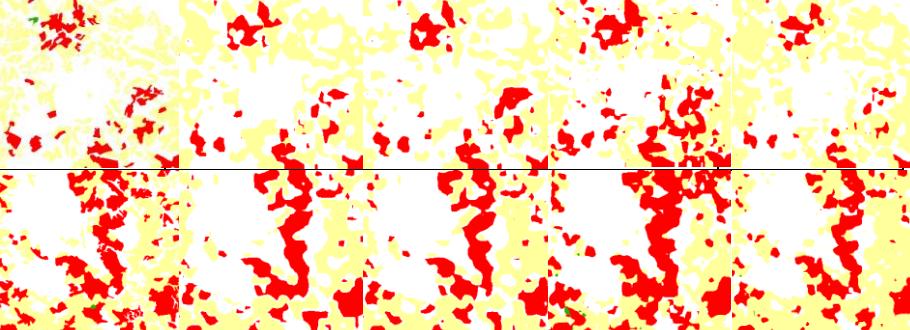
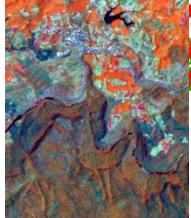
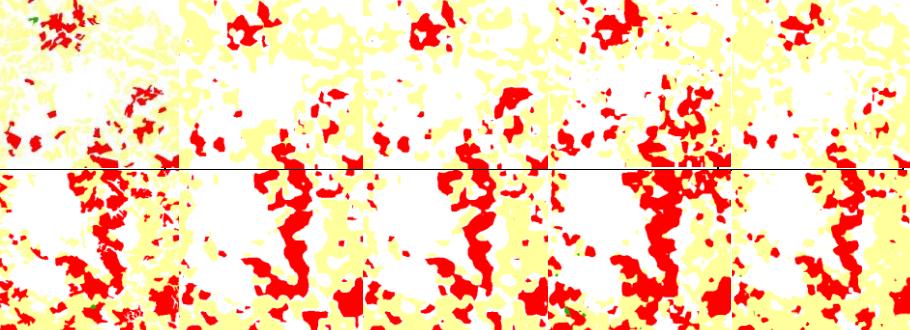
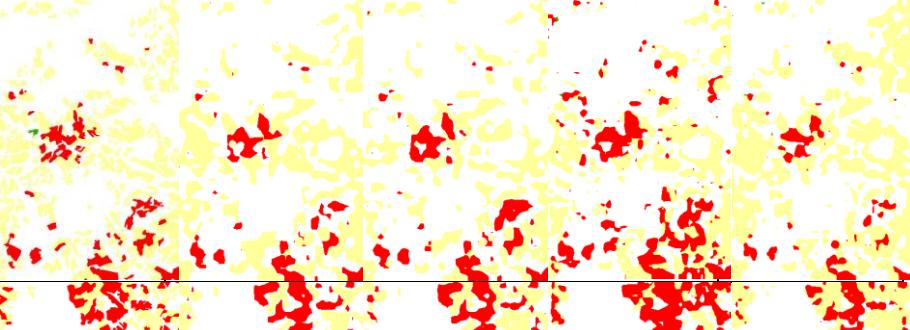
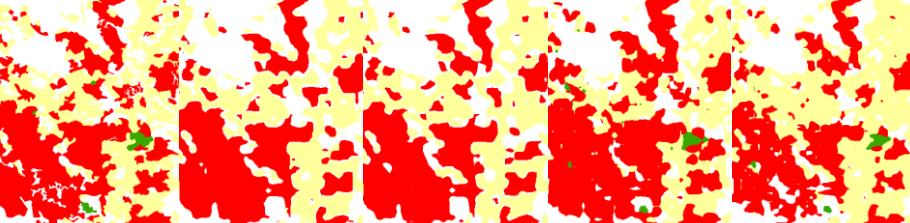
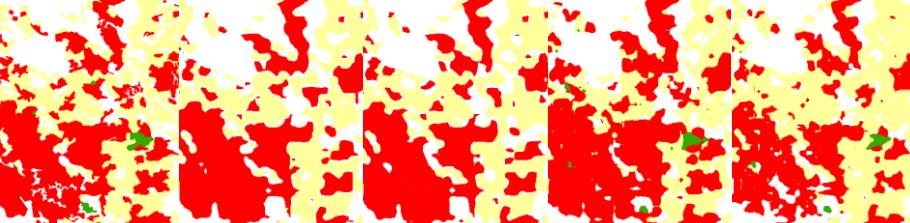
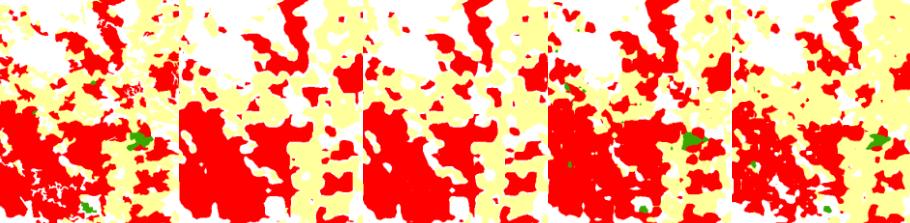
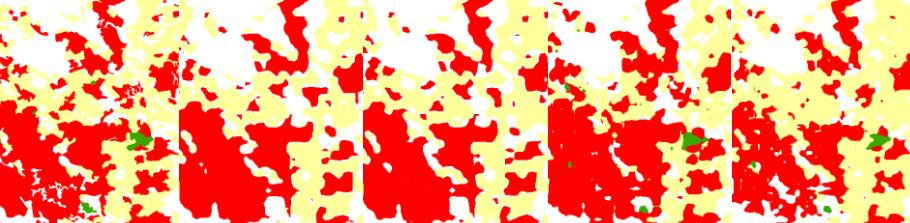
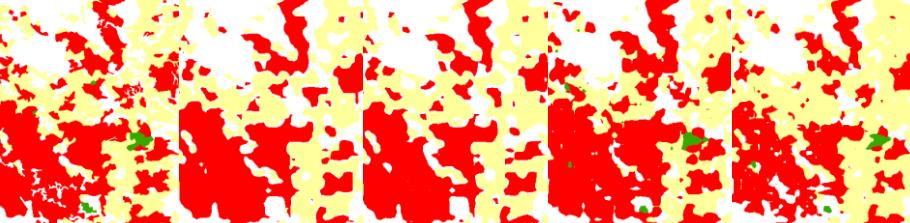
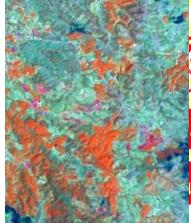
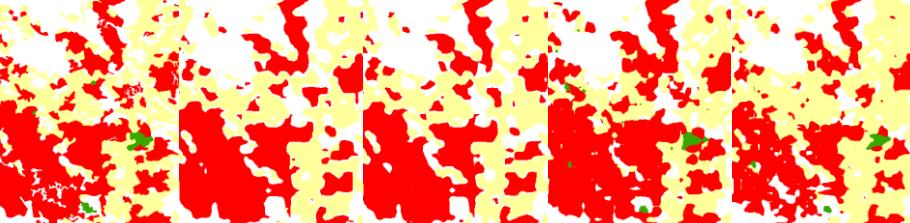
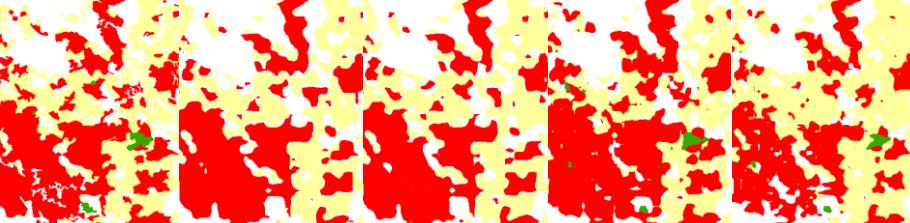
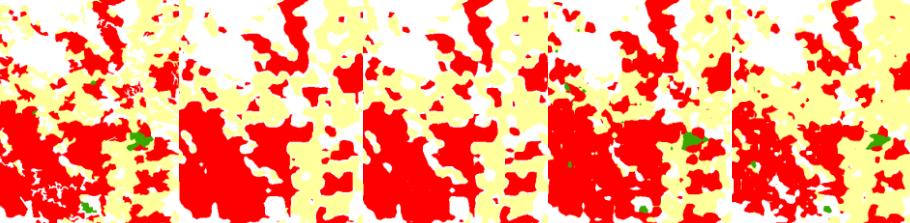
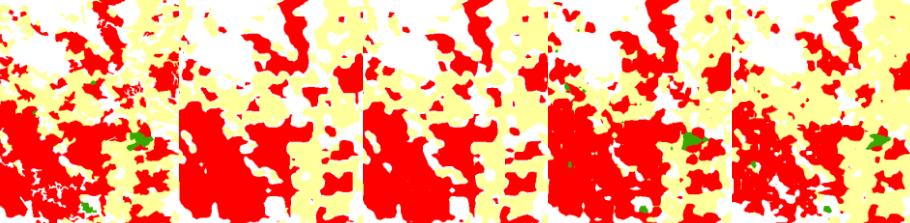
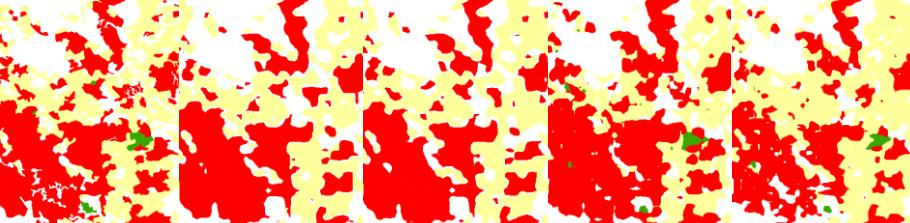
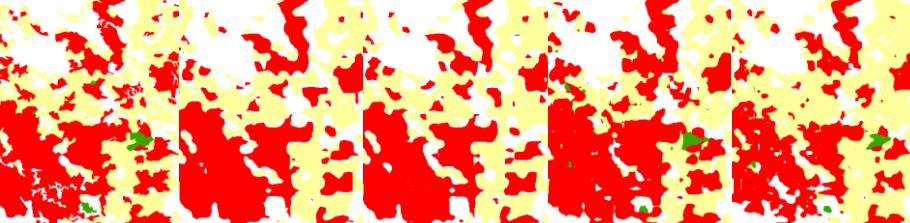
(g) Input Image	(h) Label Image	Baseline	Proposed Methods		
			(i) DCED	(j) GCN-A	(k) FF
					
					
					
					
					

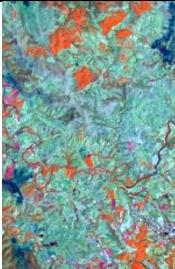
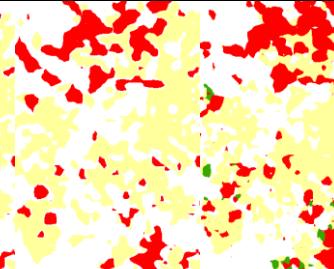
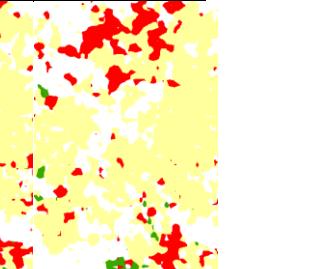
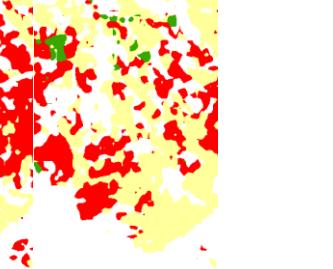
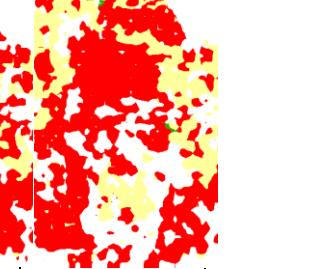
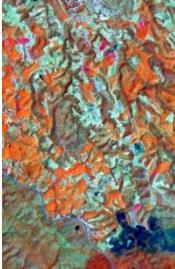
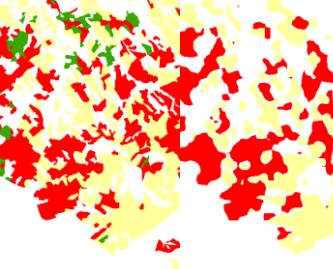
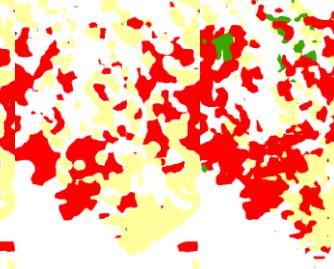
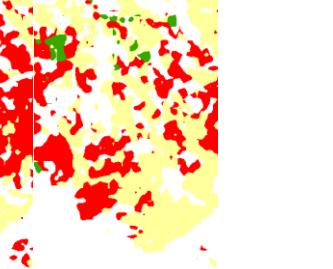
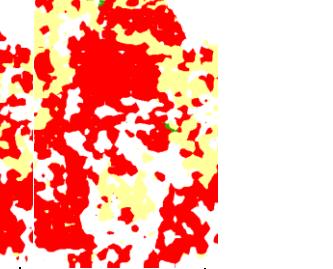
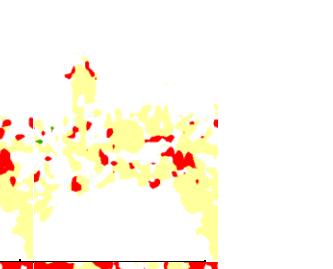
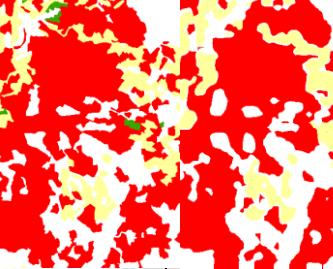
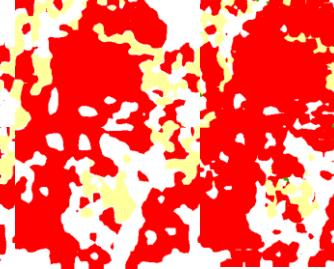
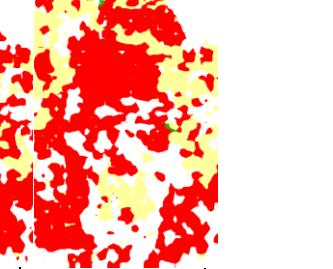
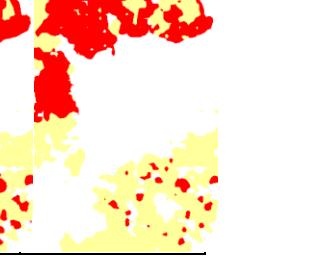
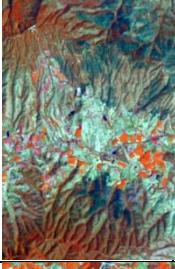
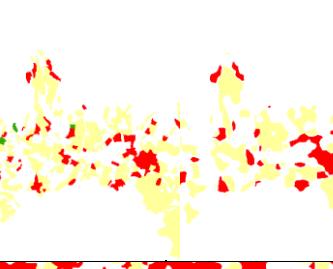
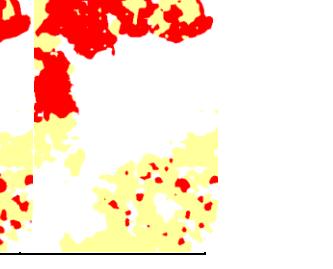
(g) Input Image	(h) Label Image	Baseline	Proposed Methods			
			(i) DCED	(j) GCN-A	(k) FF	(l) DA
						
						
						
						
						

(g) Input Image	(h) Label Image	Baseline	Proposed Methods			
			(i) DCED	(j) GCN-A	(k) FF	(l) DA
						
						
						
						
						



(g) Input Image	(h) Label Image	Baseline	Proposed Methods			
			(i) DCED	(j) GCN-A	(k) FF	(l) DA
						
						
						
						
						

(g) Input Image	(h) Label Image	Baseline	Proposed Methods			
			(i) DCED	(j) GCN-A	(k) FF	(l) DA
						
						
						
						
						

(g) Input Image	(h) Label Image	Baseline	Proposed Methods		
			(i) DCED	(j) GCN-A	(k) FF
					
					
					
					

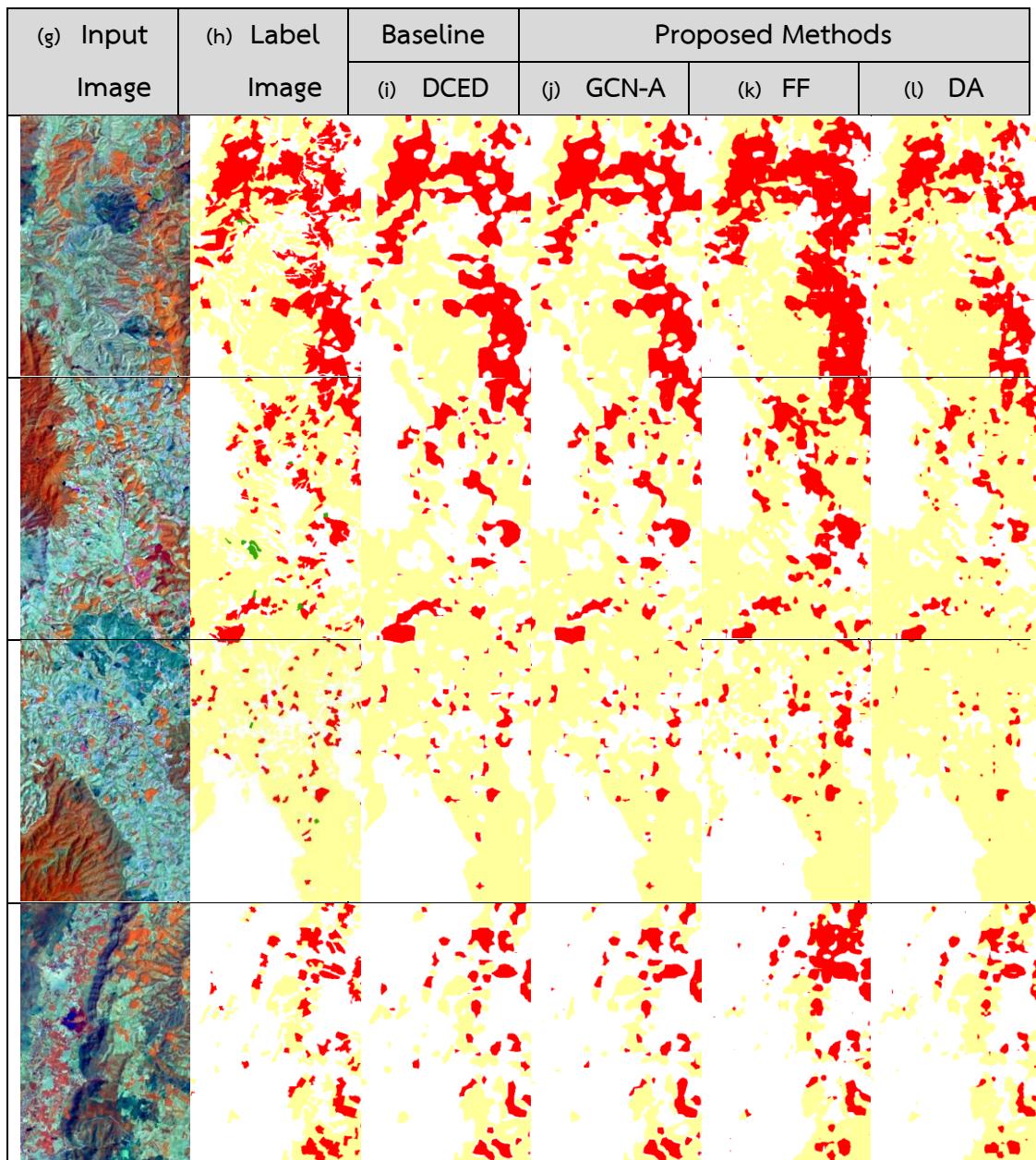


Figure 27. Testing sample input and output satellite images on Landsat-8 in Isan (Northeastern) in Thailand, where rows refer to different images. (a) Original input image. (b) Target map (ground truth). (c) Output of DCED (Baseline). (d) Output of GCN152-TL-A. (e) Output of GCN152-TL-A-FF. and (f) Output of GCN152-TL-A-FF-DA.

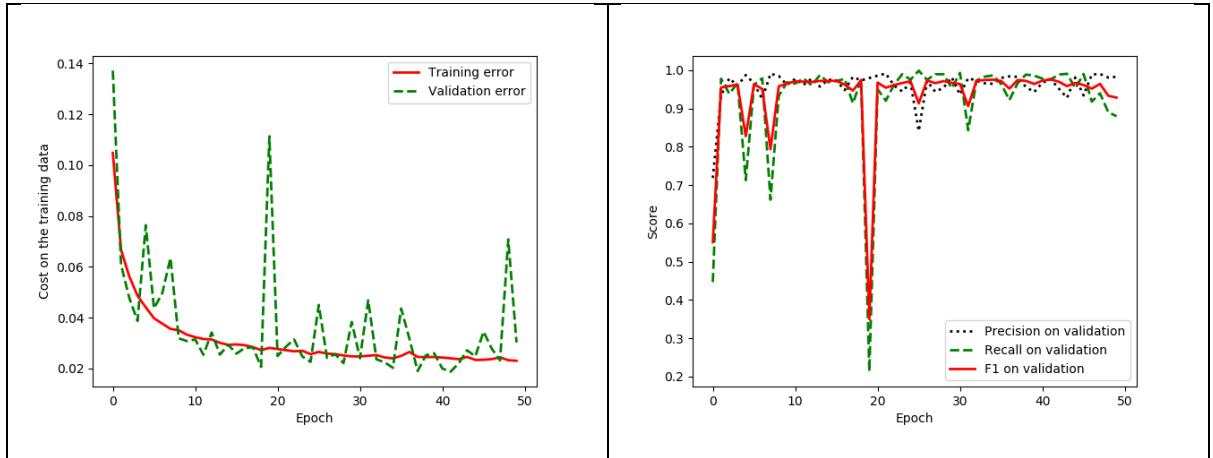


Figure 28. Graph (learning curves) on Landsat-8w3c data set of the proposed approach, “GCN152-TL-A-FF-DA”; x refers to epochs, and y refers to different measures (a) Plot of model loss (cross-entropy) on training and validation corpora; (b) performance plot on the validation corpus.

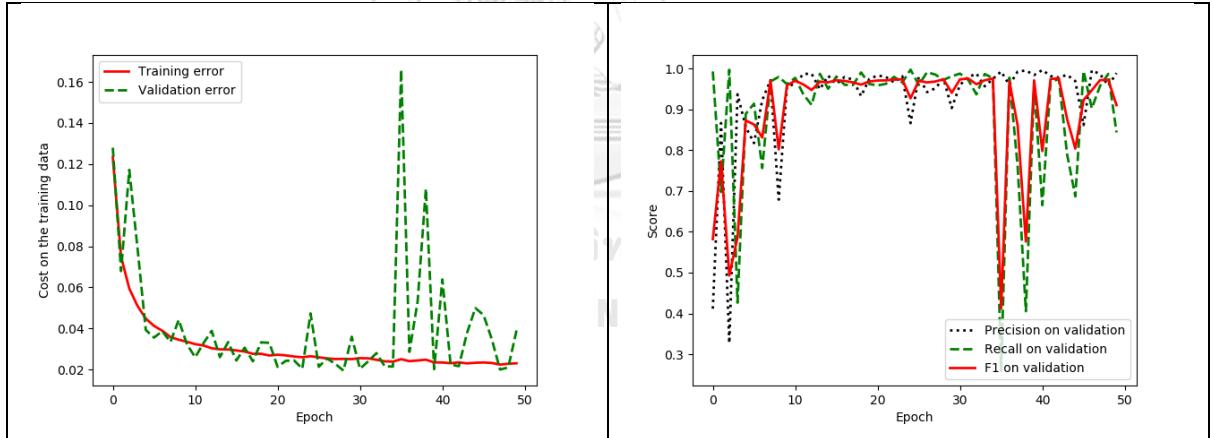


Figure 29. Graph (learning curves) on Landsat-8w3c data set of the baseline approach, DCED [1, 12, 30, 31]; x refers to epochs, and y refers to different measures (a) Plot of model loss (cross-entropy) on training and validation corpora; (b) performance plot on the validation corpus.

5.7. Results of the ISPRS Vaihingen Challenge Corpus with Discussion

An experiment was conducted on the ISPRS Vaihingen Challenge corpus, and the result is shown in Table 10 and Table 11 by comparing between baseline and variations of the proposed techniques. This shows that our network with all strategies (GCN152-TL-A-FF-DA) outperforms other methods. More details will be discussed to show that each of the proposed techniques can improve accuracy. Only in this experiment is one baseline, which is the DCED network.

Table 10. Results of the testing data of the ISPRS 2D semantic labeling challenge corpus between the baseline and five variations of our proposed techniques in terms of precision, recall, and F1-score

Method	Pretrained	Backbone	Model	Pr	Re	F1
Baseline	-	-	DCED [1]	0.867	0.849	0.854
Proposed	-	Res50	GCN	0.872	0.852	0.858
	-	Res101	GCN	0.850	0.854	0.866
	-	Res152	GCN	0.873	0.864	0.868
	-	Res152	GCN-A	0.875	0.869	0.874
	TL	Res152	GCN-A	0.897	0.877	0.881
	TL	Res152	GCN-A-FF	0.896	0.904	0.905
	TL	Res152	GCN-A-FF-DA	0.923	0.900	0.911

5.7.1. Effect of the Enhanced GCN on the ISPRS Vaihingen Corpus

Our first strategy aims to increase the F1-score of the network by varying backbones using ResNet 50, ResNet 101, and ResNet 152 rather than the traditional

one, the DCED method. From Table 10, the F1 of GCN152 (0.868) outperforms that of GCN50 (0.866), GCN101 (0.858), and the baseline method, DCED (0.854); this yields a higher F1 at 1.4%, 1.2%, and 0.4%, respectively. This can imply that an enhanced GCN is also more accurate than the DCED approach on a very high-resolution dataset. ResNet with a large number of layers is still more robust than a small number of layers, the same as that performed on the Landsat-8 corpus (Chapter 5.1 and Chapter 5.2)

When comparing the results between the original GCN method and the enhanced GCN methods on the ISPRS corpus (Table 10 and Table 11), it is clear that the GCN with a larger backbone layer can improve network performance in terms of F1-score.

Table 11. Results of the testing data of ISPRS Vaihingen Challenge corpus between each class with our proposed techniques in terms of accuracy

Method	Model	Imps	Building	Low veg	Tree	Car
Baseline	DCED [1]	0.872	0.893	0.841	0.914	0.815
Proposed	GCN50	0.876	0.873	0.857	0.953	0.803
	GCN101	0.941	0.913	0.742	0.904	0.699
	GCN152	0.810	0.963	0.895	0.912	0.806
	GCN152-A	0.886	0.928	0.811	0.895	0.820
	GCN152-TL-A	0.871	0.916	0.890	0.918	0.874
	GCN152-TL-A-FF	0.928	0.976	0.926	0.968	0.898
	GCN152-TL-A-FF-DA	0.907	0.979	0.927	0.972	0.910

5.7.2. Effect of Using Channel Attention on ISPRS Vaihingen Corpus

Our second mechanism focused on utilizing the channel attention block to change the weights of the features on each stage to enhance the consistency. From

Table 10, the F1 of GCN152-A (0.874) is greater than that of GCN152 (0.868); this yields a higher F1-score at 0.6%. The results (Figure30e) show that this can also cause the network to obtain discriminative features stage-wise to make intra-class prediction consistent with respect to very high-resolution images.

5.7.3. The Effect of Using Domain-Specific Transfer Learning on the ISPRS Corpus

This strategy aims to perform domain-specific transfer learning (details in Chapter 4.4) by reusing the pre-trained weight from the GCN152-A model on the Landsat-8 corpus. From Table, the F1 of the GCN152-TL-A method is the winner when compared with the previous proposed methods; it clearly outperforms not only the baseline but also all previous generations. Its F1 is higher than the DCED (baseline) at 2.6%. Additionally, the result illustrates that the concept of domain-specific transfer learning can enhance both precision (3%) and recall (1.8%).

5.7.4. The Effect of Using Feature Fusion on ISPRS Vaihingen Corpus

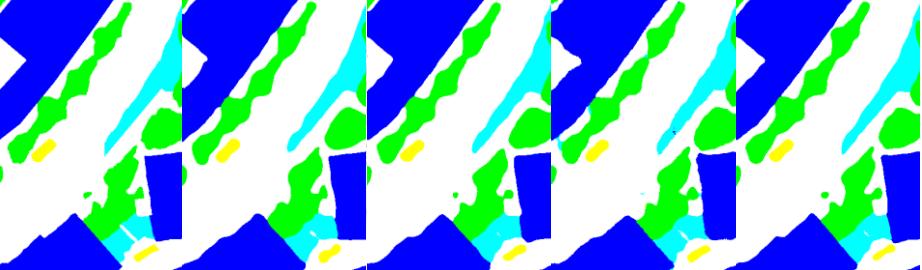
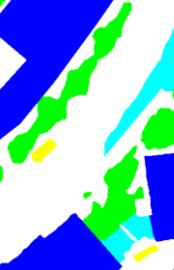
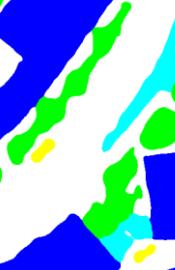
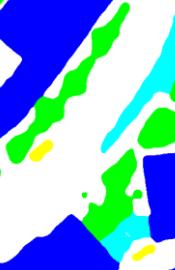
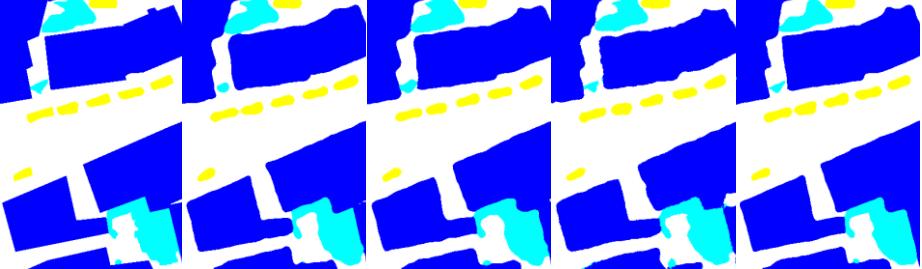
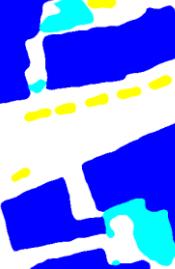
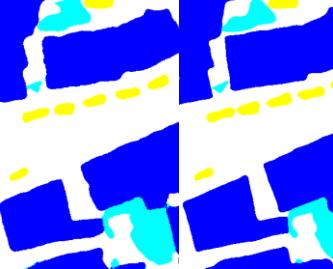
Our last strategy aims to use approach of domain-specific transfer learning (details in Chapter 4.5) by fusing two paths that different in layer of feature representation. From Table, the F1 of the GCN152-TL-A-FF method is the winner; it clearly outperforms not only the baseline but also all previous generations. Its F1 is higher than that of the DCED (baseline) at 5.1%. Additionally, the result illustrates that the concept of feature fusion, fuse the different layer of the features, can enhance both precision (2.9%) and recall (5.5%).

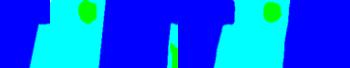
5.7.5. The Effect of Using Depthwise Atrous Convolution on on ISPRS Vaihingen Corpus

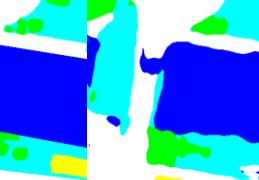
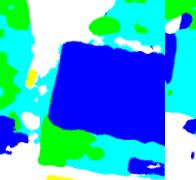
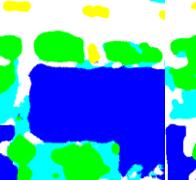
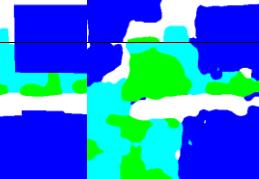
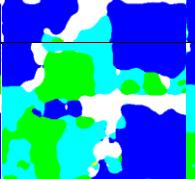
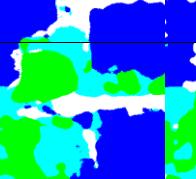
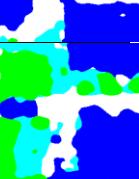
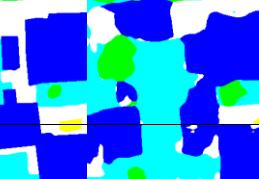
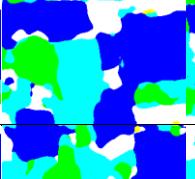
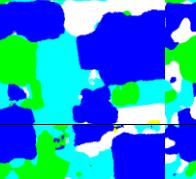
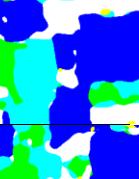
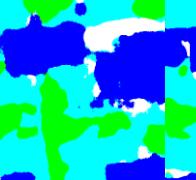
Finally, our last approach is to apply “Depthwise Atrous Convolution” to intensify the deep features from the previous step. From Tables 10 and 11 we see that the F1 of the “GCN152-TL-A-FF-DA” method is also the conqueror in this data set. The F1

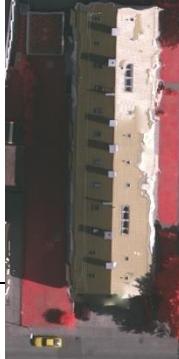
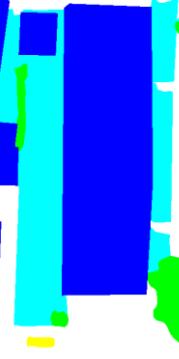
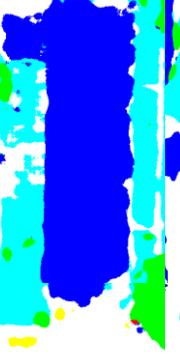
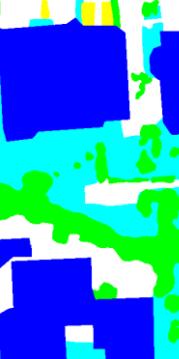
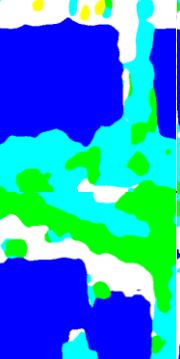
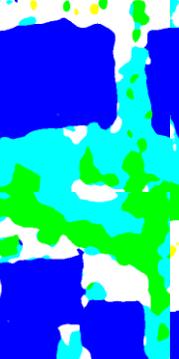
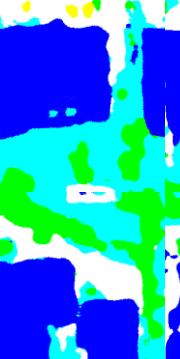
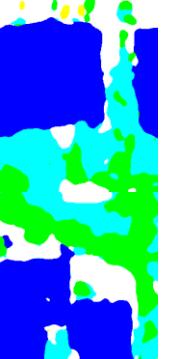
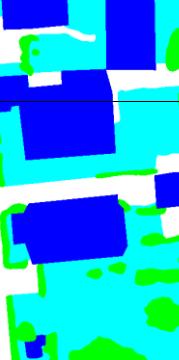
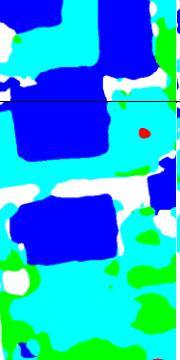
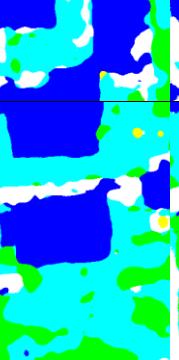
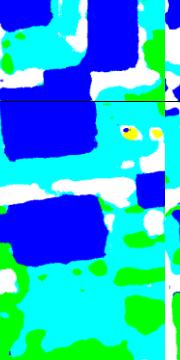
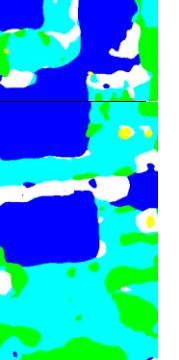
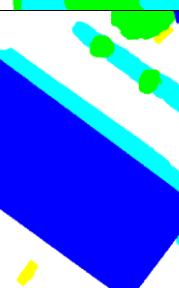
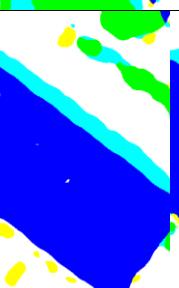
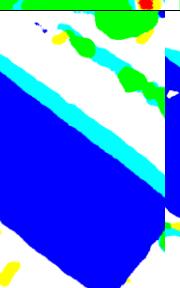
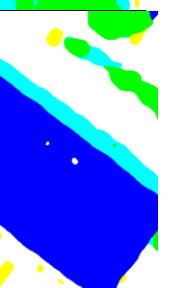
score of “GCN152-TL-A-FF-DA” is also more precise than the DCED and GCN152-TL-A-FF at 5.7% and 0.6%, consecutively.

Figure 30 shows ten sample results from the proposed method. By applying all strategies, the images in the last column (Figure23f) are similar to ground truths (Figure23b). Furthermore, F1 results is improved for each strategy we added to the network as shown in Figure30c-f and Figure30c-f.

(m) Input Image	(n) Label Image	Baseline	Proposed Methods		
		(o) DCED	(p) GCN-A	(q) FF	(r) DA
					
					
					
					

(m) Input Image	(n) Label Image	Baseline	Proposed Methods			
			(o) DCED	(p) GCN-A	(q) FF	(r) DA
						
						
						
						
						
						
						
						
						
						
						
						
						
						
						
						
						
						
						
						
						
						
						
						
						

(m) Input Image	(n) Label Image	Baseline	Proposed Methods			
			(o) DCED	(p) GCN-A	(q) FF	(r) DA
						
						
						
						
						
						
						

(m) Input Image	(n) Label Image	Baseline	Proposed Methods			
			(o) DCED	(p) GCN-A	(q) FF	(r) DA
						
						
						
						

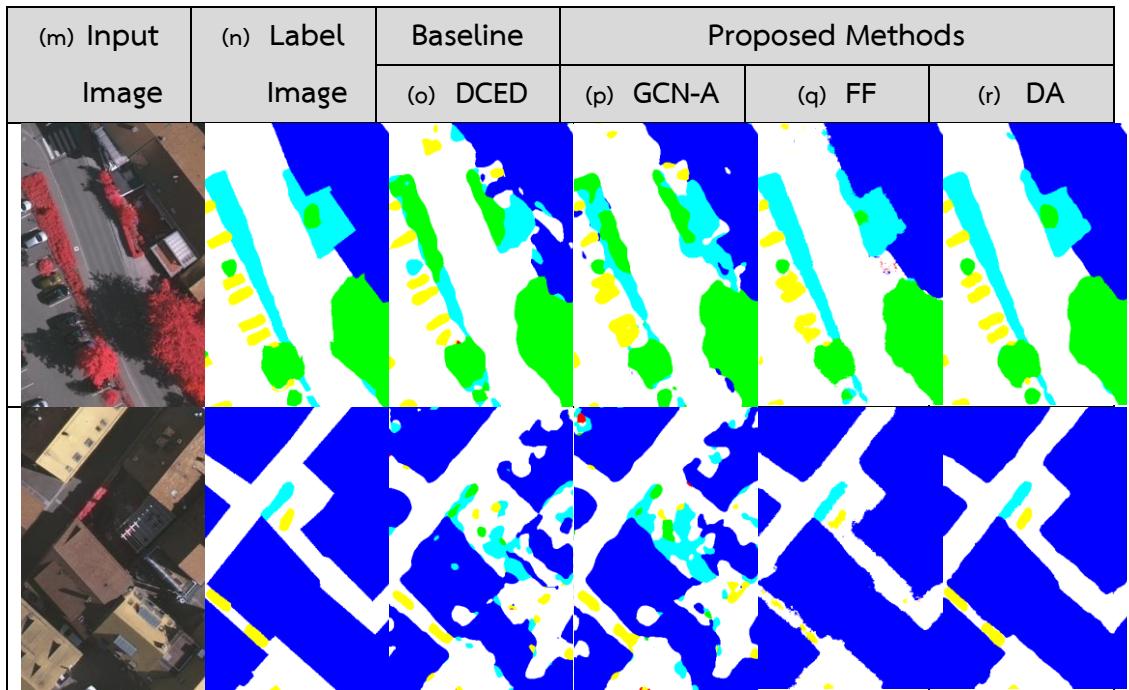


Figure 30. Comparisons between “GCN152-TL-A-FF-DA” and beyond baseline methods on the ISPRS Vaihingen (Stuttgart) challenge corpus testing set.

Figure 30 shows twenty sample testing results from the proposed method on ISPRS Vaihingen corpus. The results of the last column are also similar to the ground truth in the second column same as performed on Landsat-8 corpus. Considering to each class (are shown in Table 10 and Table 11), every class from our proposed methods are the winner in term Accuracy.

All extensive experiments on the Landsat-8 and ISPRS datasets demonstrate that the proposed method clearly achieves promising gains compared with the baseline approach.

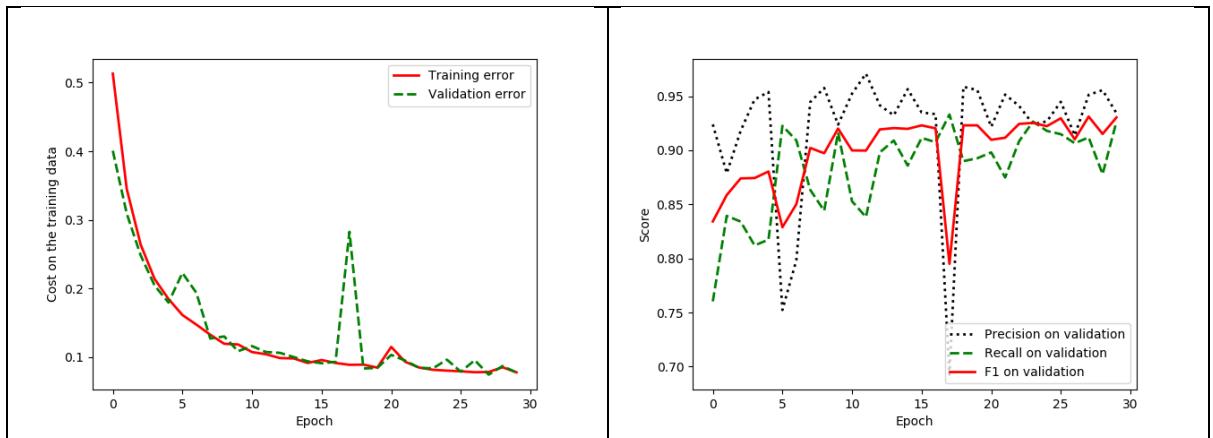


Figure 31. Graph (learning curves) on ISPRS Vaihingen Challenge corpus data set of the proposed approach, “GCN152-TL-A-FF-DA”; x refers to epochs, and y refers to different measures (a) Plot of model loss (cross-entropy) on training and validation corpora; (b) performance plot on the validation corpus.

As can be seen in Figure 30, the performance of our best model outperforms other advanced models by a considerable margin on each category, especially for the impervious surface (Imps), tree, and car categories. To show the effectiveness of the proposed methods, we performed comparisons against a number of state-of-the-art semantic segmentation methods, as listed in Table 9 and Table 10 with respect to the ISPRS corpus, and Table 5 to Table 8 with respect to the Landsat-8 corpus. The DCED [12, 30, 31] and GCN [33] are the versions with ResNet-50 as their backbone. In particular, we re-implemented the DCED with Tensorflow-Slim [76], since the released code was built on Caffe [77]. We can see that our proposed methods significantly outperform other methods on both the F1-score. Furthermore, the curves shown in Figure 31a-b exhibit that our model performs better on all given categories (Figure 32a-b).

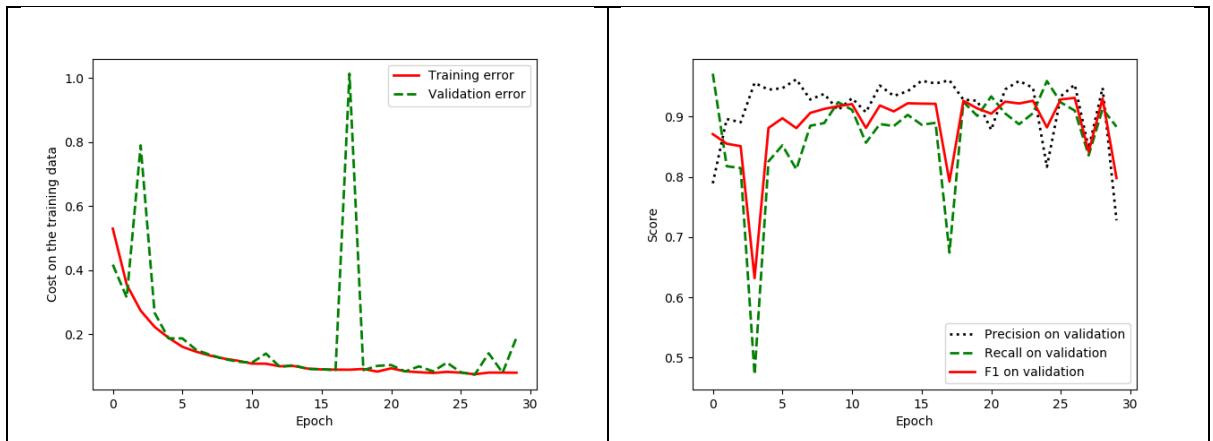


Figure 32. Graph (learning curves) on ISPRS Vaihingen Challenge corpus of the baseline approach, DCED [1, 12, 30, 31]; x refers to epochs, and y refers to different measures (a) Plot of model loss (cross-entropy) on training and validation corpora; (b) performance plot on the validation corpus.

In terms of the computational cost, our framework requires slightly additional training time compared to the baseline approach, DCED, by about 6.25% (6–7 hours), and GCN, by about 4.5% (4–5 hour). In our experiment, DCED’s training procedure took approximately 16 hour per dataset, and finished after 50 epochs with 1,152 second per epoch. Our framework is a modification of the GCN-based deep learning architecture. The channel attention model increases the time by 20 min compared with the GCN152 method and feature fusion model increases the time by 15 min compared with the GCN152-TL-A. There is no additional time required when reusing pre-trained weights.

Moreover, there are many experiments before it reached the final proposed method. These are the lists of the whole of experiments that perform not well on the Landsat-8 Satellite and the ISPRS Vaihingen Challenge Corpora.

- (1) Encoder-Decoder based on SegNet [12, 30]. We use a VGG-style encoder-decoder, where the upsampling in the decoder is done using transposed convolutions.
- (2) Encoder-Decoder based on UNet [32]. We use a U-Shape style encoder-decoder, where the upsampling in the decoder is done using transposed convolutions. Also, it employs additive skip connections from the encoder to the decoder.
- (3) The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation [24]. We use a U-shape style of encoder-decoder architecture. Each stage between the pooling layers uses dense blocks. Besides, it concatenated skip connections from the encoder to the decoder.
- (4) DeepLabV3 network [25]. We employ Atrous Spatial Pyramid Pooling to obtain multi-scale context by using multiple atrous rates.
- (5) Dense Decoder Shortcut Connections for Single-Pass Semantic Segmentation [16]. We utilize dense connectivity in the decoder step of the segmentation model.
- (6) BiSeNet: Bilateral Segmentation Network for Real-time Semantic Segmentation [14]. We apply a Spatial Path with a short stride to conserve the spatial information and produce high-resolution features while having a parallel Context Path with a speedy downsampling approach to reach a sufficient receptive field.
- (7) Pyramid Scene Parsing Network [18]. We use capability of global context information by different-region based context aggregation is applied through a pyramid pooling module.
- (8) Full-Resolution Residual Networks for Semantic Segmentation in Street Scenes [78]. We combine a multi-scale context with pixel-level accuracy by using two processing streams within the architecture. The residual stream offers information at the full image resolution, enabling precise adherence to segment boundaries.

CHAPTER VI

CONCLUSIONS

In this dissertation, we propose a novel CNN framework to perform semantic labeling and achieve image labeling on remotely sensed images. Our proposed method delivers excellent performance by presenting five aspects. (i) A global convolutional network (GCN) is employed and enhanced by adding larger numbers of layers to capture sophisticated features better. (ii) Channel attention is proposed to assign a proper weight for each extracted feature on different stages of the network. (iii) domain-specific transfer learning is introduced to allay the scarcity issue by training the initial weights using other remotely sensed corpora whose resolutions can be different. (vi) applying the "Feature Fusion (FF)" for capturing low-level features, and the last, (v) using the concept of "Depthwise Atrous Convolution (DA)" for refining the features and provide more coverage areas. The experiments were conducted on three data sets: Landsat-8w3c, Landsat-8w5c corpora (medium resolution), and the ISPRS Vaihingen Challenge (very high resolution) corpus. The results show that our model that combines all proposed strategies outperforms baseline models in terms of F1 score. The final results show that our "GCN152-TL-A-FF-DA" model outperforms the baseline (DCED)—5.7% for F1 on the Landsat-8w3c, 10.1% for F1 on the Landsat-8w5c, and 6.2% on the ISPRS corpus. Moreover, it reaches an accuracy surpassing 90% in almost all classes.

In the future, more semantic labeling task, modern optimization techniques, and other novel activation functions will be investigated and compared to obtain the best GCN-based framework for semantic segmentation in remotely sensed images. Moreover, incorporating additional data sources (e.g., a digital surface model) might be needed to increase the accuracy of deep learning for both the CNN and the modern

deep learning layer with overconfident predictions simultaneously. These issues mentioned above will be investigated in future research.



REFERENCES

1. Liu, Y., et al., *Semantic labeling in very high resolution images via a self-cascaded convolutional neural network*. ISPRS Journal of Photogrammetry and Remote Sensing, 2018. **145**: p. 78-95.
2. Wang, H., et al., *Gated convolutional neural network for semantic segmentation in high-resolution images*. Remote Sensing, 2017. **9**(5): p. 446.
3. Zhu, X.X., et al., *Deep learning in remote sensing: A comprehensive review and list of resources*. IEEE Geoscience and Remote Sensing Magazine, 2017. **5**(4): p. 8-36.
4. Yi, Y., et al., *Semantic Segmentation of Urban Buildings from VHR Remote Sensing Imagery Using a Deep Convolutional Neural Network*. Remote Sensing, 2019. **11**(15): p. 1774.
5. Marcos, D., et al., *Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models*. ISPRS journal of photogrammetry and remote sensing, 2018. **145**: p. 96-107.
6. Ma, L., et al., *Deep learning in remote sensing applications: A meta-analysis and review*. ISPRS journal of photogrammetry and remote sensing, 2019. **152**: p. 166-177.
7. Mnih, V., *Machine learning for aerial image labeling*. 2013: Citeseer.
8. Muruganandham, S., *Semantic segmentation of satellite images using deep learning*. 2016.
9. Li, Y., et al. *Fully convolutional instance-aware semantic segmentation*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.
10. Vu, T.-H., et al., *DADA: Depth-aware Domain Adaptation in Semantic Segmentation*. arXiv preprint arXiv:1904.01886, 2019.
11. Chen, L.-C., et al., *Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs*. IEEE

- transactions on pattern analysis and machine intelligence, 2017. **40**(4): p. 834-848.
12. Badrinarayanan, V., A. Kendall, and R. Cipolla, *Segnet: A deep convolutional encoder-decoder architecture for image segmentation*. IEEE transactions on pattern analysis and machine intelligence, 2017. **39**(12): p. 2481-2495.
 13. Chen, L.-C., et al. *Encoder-decoder with atrous separable convolution for semantic image segmentation*. in *Proceedings of the European conference on computer vision (ECCV)*. 2018.
 14. Yu, C., et al. *Bisenet: Bilateral segmentation network for real-time semantic segmentation*. in *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.
 15. Zhang, H., et al. *Context encoding for semantic segmentation*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.
 16. Bilinski, P. and V. Prisacariu. *Dense decoder shortcut connections for single-pass semantic segmentation*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.
 17. Yang, M., et al. *Denseaspp for semantic segmentation in street scenes*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.
 18. Zhao, H., et al. *Pyramid scene parsing network*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
 19. Tian, Z., et al. *Decoders Matter for Semantic Segmentation: Data-Dependent Decoding Enables Flexible Feature Aggregation*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.
 20. Poudel, R.P., S. Liwicki, and R. Cipolla, *Fast-SCNN: fast semantic segmentation network*. arXiv preprint arXiv:1902.04502, 2019.
 21. Li, X., et al., *GFF: Gated Fully Fusion for Semantic Segmentation*. arXiv preprint arXiv:1904.01803, 2019.
 22. Islam, M.A., et al., *Gated feedback refinement network for coarse-to-fine dense semantic image labeling*. arXiv preprint arXiv:1806.11266, 2018.

23. Ziegler, T., et al., *Efficient Smoothing of Dilated Convolutions for Image Segmentation*. arXiv preprint arXiv:1903.07992, 2019.
24. Jégou, S., et al. *The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2017.
25. Chen, L.-C., et al., *Rethinking atrous convolution for semantic image segmentation*. arXiv preprint arXiv:1706.05587, 2017.
26. Panboonyuen, T., et al., *Road segmentation of remotely-sensed images using deep convolutional neural networks with landscape metrics and conditional random fields*. Remote Sensing, 2017. **9**(7): p. 680.
27. He, K., et al. *Mask r-cnn*. in *Proceedings of the IEEE international conference on computer vision*. 2017.
28. Panboonyuen, T., et al. *An enhanced deep convolutional encoder-decoder network for road segmentation on aerial imagery*. in *International Conference on Computing and Information Technology*. 2017. Springer.
29. Panboonyuen, T., et al., *Semantic segmentation on remotely sensed images using an enhanced global convolutional network with channel attention and domain specific transfer learning*. Remote Sensing, 2019. **11**(1): p. 83.
30. Badrinarayanan, V., A. Handa, and R. Cipolla, *Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling*. arXiv preprint arXiv:1505.07293, 2015.
31. Kendall, A., V. Badrinarayanan, and R. Cipolla, *Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding*. arXiv preprint arXiv:1511.02680, 2015.
32. Ronneberger, O., P. Fischer, and T. Brox. *U-net: Convolutional networks for biomedical image segmentation*. in *International Conference on Medical image computing and computer-assisted intervention*. 2015. Springer.
33. Peng, C., et al. *Large Kernel Matters--Improve Semantic Segmentation by Global Convolutional Network*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

34. Panboonyuen, T., et al., *Semantic segmentation on remotely sensed images using an enhanced global convolutional network with channel attention and domain specific transfer learning*. 2019. **11**(1): p. 83.
35. Yu, C., et al. *Learning a discriminative feature network for semantic segmentation*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.
36. Hu, J., L. Shen, and G. Sun. *Squeeze-and-excitation networks*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.
37. Xie, M., et al. *Transfer learning from deep features for remote sensing and poverty mapping*. in *Thirtieth AAAI Conference on Artificial Intelligence*. 2016.
38. Yosinski, J., et al. *How transferable are features in deep neural networks?* in *Advances in neural information processing systems*. 2014.
39. Liu, J., Y. Wang, and Y. Qiao. *Sparse deep transfer learning for convolutional neural network*. in *Thirty-First AAAI Conference on Artificial Intelligence*. 2017.
40. Vieira, S., et al., *Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications*. 2017. **74**: p. 58-75.
41. Hidaka, A. and T. Kurita. *Consecutive dimensionality reduction by canonical correlation analysis for visualization of convolutional neural networks*. in *Proceedings of the ISCIE International Symposium on Stochastic Systems Theory and its Applications*. 2017. The ISCIE Symposium on Stochastic Systems Theory and Its Applications.
42. Snuverink, I., *Deep Learning for Pixelwise Classification of Hyperspectral Images: A generalizing model for a fixed scene subject to temporally changing weather, lighting and seasonal conditions*. 2017.
43. Noh, H., S. Hong, and B. Han. *Learning deconvolution network for semantic segmentation*. in *Proceedings of the IEEE international conference on computer vision*. 2015.
44. Long, J., E. Shelhamer, and T. Darrell. *Fully convolutional networks for semantic segmentation*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

45. Audebert, N., B. Le Saux, and S. Lefèvre. *Semantic segmentation of earth observation data using multimodal and multi-scale deep networks*. in *Asian conference on computer vision*. 2016. Springer.
46. Valada, A., et al. *Adapnet: Adaptive semantic segmentation in adverse environmental conditions*. in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. 2017. IEEE.
47. Zhao, H., et al. *Icnet for real-time semantic segmentation on high-resolution images*. in *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.
48. Dai, J., K. He, and J. Sun. *Instance-aware semantic segmentation via multi-task network cascades*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.
49. Tian, Z., et al., *FCOS: Fully Convolutional One-Stage Object Detection*. arXiv preprint arXiv:1904.01355, 2019.
50. Zhou, Y., et al. *Context-Reinforced Semantic Segmentation*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.
51. Wu, T., et al., *Tree-structured Kronecker Convolutional Networks for Semantic Segmentation*. arXiv preprint arXiv:1812.04945, 2018.
52. Zhao, H., et al. *Psanet: Point-wise spatial attention network for scene parsing*. in *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.
53. Song, C., et al. *Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.
54. He, T., et al. *Knowledge Adaptation for Efficient Semantic Segmentation*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.
55. Ding, H., et al. *Context contrasted feature and gated multi-scale aggregation for scene segmentation*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

56. Liu, C., et al. *Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.
57. Liu, Y., et al. *Structured Knowledge Distillation for Semantic Segmentation*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.
58. He, K., et al. *Deep residual learning for image recognition*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
59. Ioffe, S. and C. Szegedy, *Batch normalization: Accelerating deep network training by reducing internal covariate shift*. arXiv preprint arXiv:1502.03167, 2015.
60. Kingma, D.P. and J. Ba, *Adam: A method for stochastic optimization*. arXiv preprint arXiv:1412.6980, 2014.
61. Liu, X., Z. Deng, and Y. Yang, *Recent progress in semantic image segmentation*. Artificial Intelligence Review, 2019. **52**(2): p. 1089-1106.
62. Tan, M. and Q.V. Le, *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. arXiv preprint arXiv:1905.11946, 2019.
63. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning*. nature, 2015. **521**(7553): p. 436.
64. Garcia-Garcia, A., et al., *A review on deep learning techniques applied to semantic segmentation*. arXiv preprint arXiv:1704.06857, 2017.
65. Thoma, M., *A survey of semantic segmentation*. arXiv preprint arXiv:1602.06541, 2016.
66. Kampffmeyer, M., R. Jenssen, and A.-B. Salberg. *Dense Dilated Convolutions Merging Network for Semantic Mapping of Remote Sensing Images*. in *2019 Joint Urban Remote Sensing Event (JURSE)*. 2019. IEEE.
67. Pang, Y., et al. *Towards bridging semantic gap to improve semantic segmentation*. in *Proceedings of the IEEE International Conference on Computer Vision*. 2019.
68. Diakogiannis, F.I., et al., *ResUNet-a: a deep learning framework for semantic segmentation of remotely sensed data*. arXiv preprint arXiv:1904.00592, 2019.

69. Zhang, Z., et al. *Exfuse: Enhancing feature fusion for semantic segmentation*. in *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.
70. Yang, W., et al., *Lightweight feature fusion network for single image super-resolution*. 2019. **26**(4): p. 538-542.
71. Ma, C., X. Mu, and D.J.I.A. Sha, *Multi-layers feature fusion of convolutional neural network for scene classification of remote sensing*. 2019. **7**: p. 121685-121694.
72. Du, Y., et al., *Deep learning with multi-scale feature fusion in remote sensing for automatic oceanic eddy detection*. 2019. **49**: p. 89-99.
73. Duarte, D., et al., *Multi-resolution feature fusion for image classification of building damages with convolutional neural networks*. 2018. **10**(10): p. 1636.
74. Panboonyuen, T., et al., *Semantic Labeling in Remote Sensing Corpora Using Feature Fusion-Based Enhanced Global Convolutional Network with High-Resolution Representations and Depthwise Atrous Convolution*. 2020. **12**(8): p. 1233.
75. Ghosh, S., et al., *Understanding deep learning techniques for image segmentation*. 2019. **52**(4): p. 1-35.
76. Abadi, M., et al. *Tensorflow: A system for large-scale machine learning*. in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*. 2016.
77. Jia, Y., et al. *Caffe: Convolutional architecture for fast feature embedding*. in *Proceedings of the 22nd ACM international conference on Multimedia*. 2014. ACM.
78. Pohlen, T., et al. *Full-resolution residual networks for semantic segmentation in street scenes*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.



จุฬาลงกรณ์มหาวิทยาลัย

CHULALONGKORN UNIVERSITY

VITA

NAME	ธีรพงศ์ ปานบุญยืน (Teerapong Panboonyuen)
DATE OF BIRTH	5 December 1992
PLACE OF BIRTH	Ratchaburi, Thailand
INSTITUTIONS ATTENDED	Chulalongkorn University
HOME ADDRESS	433/54 Srisuriyawong Rd., Na-muang, Muang, Ratchaburi, 70000
PUBLICATION	[1] Panboonyuen, T.; Jitkajornwanich, K.; Lawawirojwong, S.; Srestasathiern, P.; Vateekul, P. Semantic Labeling in Remote Sensing Corpora Using Feature Fusion-Based Enhanced Global Convolutional Network with High-Resolution Representations and Depthwise Atrous Convolution. <i>Remote Sens.</i> 2020, 12, 1233. [2] Panboonyuen, T.; Jitkajornwanich, K.; Lawawirojwong, S.; Srestasathiern, P.; Vateekul, P. Semantic Segmentation on Remotely Sensed Images Using an Enhanced Global Convolutional Network with Channel Attention and Domain Specific Transfer Learning. <i>Remote Sens.</i> 2019, 11, 83. [3] Panboonyuen, T.; Jitkajornwanich, K.; Lawawirojwong, S.; Srestasathiern, P.; Vateekul, P. Road Segmentation of Remotely-Sensed Images Using Deep Convolutional Neural Networks with Landscape Metrics and Conditional Random Fields. <i>Remote Sens.</i> 2017, 9, 680. [4] Panboonyuen, T., Vateekul, P., Jitkajornwanich, K., & Lawawirojwong, S. (2017, July). An enhanced deep convolutional encoder-decoder network for road segmentation on aerial imagery. In International

Conference on Computing and Information Technology (pp. 191-201). Springer, Cham.

[5] Panboonyuen, Teerapong, et al. "Image Vectorization of Road Satellite Data Sets", Journal of Remote Sensing and GIS Association of Thailand (2017)

[6] Wichakam, I., Panboonyuen, T., Udomcharoenchaikit, C., & Vateekul, P. (2018, February). Real-Time Polyps Segmentation for Colonoscopy Video Frames Using Compressed Fully Convolutional Network. In International Conference on Multimedia Modeling (pp. 393-404). Springer, Cham.

[7] Jitkajornwanich, K., Vateekul, P., Panboonyuen, T., Lawavirojwong, S., & Srisonphan, S. (2017, December).

Road map extraction from satellite imagery using connected component analysis and landscape metrics. In 2017 IEEE International Conference on Big Data (Big Data) (pp. 3435-3442). IEEE.

[8] Chantharaj, S., Pornratthanapong, K., Chitsinpchayakun, P., Panboonyuen, T., Vateekul, P., Lawavirojwong, S., ... & Jitkajornwanich, K. (2018, July). Semantic Segmentation On Medium-Resolution Satellite Images Using Deep Convolutional Networks With Remote Sensing Derived Indices. In 2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE) (pp. 1-6). IEEE.

[9] Kantavat, P., Hayashi, Y., City, G. S., Kijsirikul, B., Panboonyuen, T., Achariyaviriya, W., ... & Vateekul, P. Transportation Mobility Factor Extraction Using Image Recognition Techniques, First International Conference on Smart Technology & Urban Development (STUD 2019).

AWARD RECEIVED**AWARDS AND HONORS**

- H.M. the King Bhumibol Adulyadej's 72nd Birthday Anniversary Scholarship (Master Degree)
- The 100th Anniversary Chulalongkorn University Fund for Doctoral Scholarship (Ph.D.)
- The 90th Anniversary Chulalongkorn University Fund (Ratchadaphiseksomphot Endowment Fund)
- Best Student Paper Award in International Conference on Computing and Information Technology 2017
- Web of Science ResearcherID: AAO-4985-2020 (<https://publons.com/researcher/AAO-4985-2020/>)
- Google Scholar (<https://scholar.google.co.th/citations?user=myy0qDgAAAAJ&hl=en>)

PROFESSIONAL SERVICE

- Invited Reviewer of International Journal of Remote Sensing (Tier1, Q1, ISI Journal)
- Invited Reviewer of Sensors (Q2, Journal)
- Invited Reviewer of IEEE Transactions on Industrial Informatics (Q1, ISI Journal)
- Invited Reviewer of IEEE Access (Tier1, Q1, ISI Journal)
- Invited Reviewer of IEEE Transactions on Geoscience and Remote Sensing (Tier1, Q1, ISI Journal)