東京大学大学院新領域創成科学研究科
社会文化環境学専攻

2020 年度

修士論文

深層学習に基づく航空写真とデジタル地表モデルのエンドツーエンドの建物変化検出モデル

End-to-end Building Change Detection Model in Aerial Imagery and Digital Surface Model Based on Neural Networks

2020 年 7 月 10 日提出
指導教員　柴崎　亮介　教授

連　欣蕾
Lian, Xinlei

Abstract

Multi-temporal building change detection is one of the most essential major issues of photogrammetry and remote sensing at current stage, which is of great significance for wide applications as offering real estate indicators as well as monitoring urban environment. Although current photogrammetry methodologies could be applied to 2-D remote sensing imagery for rectification with sensor parameters, multi-temporal aerial or satellite imagery is not adequate to offer spectral and textual features for building change detection. Alongside recent development of Dense Image Matching (DIM) technology, the acquisition of 3-D point cloud and Digital Surface Model (DSM) has been generally realized, which could be combined with imagery, making building change detection more effective with greater spatial structure and texture information. Over the past years, scholars have put forward vast change detection techniques including traditional and model-based solutions. Nevertheless, existing appropriate methodology combined with Neural Networks (NN) for accurate building change detection with multi-temporal imagery and DSM remains to be of great research focus currently due to the inevitable limitations and omissions of existing NN-based methods, which is of great research prospect. This study proposed a novel end-to-end model framework based on deep learning for pixel-level building change detection from high-spatial resolution aerial ortho imagery and corresponding DSM sharing same resolution, which is from the dataset of Tokyo whole area.

Key Words: Building Change Detection, End-To-End, Feature Pyramid Network (FPN), digital surface model (DSM), Change Map

Acknowledgements

Foremost, I would like to deliver my sincere and deepest gratitude to my supervisors Prof. Shibasaki, Prof. Song and Prof. Shao. From the bottom of my heart, I would like to appreciate my honorable adviser, Prof. Shibasaki. Prof. Shibasaki has been continually supporting my academic life including my master research and study, as well as improving myself. He accepted me to Shibasaki. Lab two years ago, letting me open a new chapter of new life in University of Tokyo and Japan. During my master course, he has always giving me valuable suggestions after listening to my research presentation carefully. Besides, he offered me opportunities to join projects cooperating with corporations, and to present in meeting of Center of Spatial Information Sciences (CSIS), which let me enhance myself in various aspects. I would like to specially thank Prof. Song, who is the leader of Intelligent Perception and urban Computing (IPUC) group which I affiliated to. He has been giving me biggest understanding and support throughout my master years. My special sincere thank also goes to Prof. Shao, for his instructions to me on methodologies of computer vision and remote sensing.

In addition, I am grateful for my lab mates in Shibasaki Lab: thanks to researcher Dr. Yuan Wei, who has given me adequate instructions on remote sensing and prepared auxiliary data for my research, contributing much on my master thesis research. Thanks to my senior student Dr. Guo Zhiling, who has great experience in computer vision and deep learning, giving me valid resolutions whenever my research was stuck. Besides, the thank also goes to my peer master student and good friend Cai Zekun, who is excellent in Computer Science and Programming, having given me countless help throughout my whole research.

Also, I would like to appreciate other researchers and senior students Zhang Haoran, Jiang Renhe, Chen Quanjun, Huang Dou, Shi Xiaodan, Song Xiaoya...who I had the opportunities to work, study or discuss together. They are always kind and enthusiastic to help me, promoting me to get used to life in Japan and Shibasaki. Lab. For my peer students, Cai Zekun, Chen Pochia, Yao Yuhao, Li Wenjing and Chen Jinyu, who are all my great friends. We spent most of the master years together, taking master entrance exam, taking courses, preparing for group meeting and hanging out. They are the warmest and most sparkling part of my U-Tokyo master life.

I would like to acknowledge the University of Tokyo and the Department of my Socio-Cultural Environmental Studies who gave me precious study life. Another acknowledgement goes to NTT Spatial Information that provided high-quality data for my master research.

Finally, I would like to appreciate all my dear friends both in Japan and other countries, my family and my beloved one who have always supported me during these two years.

Contents

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

List of Tables

Abbreviations

DIM	***D**ense **I**mage **M**atching*
DSM	***D**igital **S**urface **M**odel*
NDVI	***N**ormalized **D**ifference **V**egetation **I**ndex*
FPN	***F**eature **P**yramid **N**etwork*
CNN	***C**onvolutional **N**eural **N**etwork*
SDAE	***S**tacked **D**enoising **A**utoencoders*
SCV	***S**emi-supervised **C**han–**V**ese*
HSI-CD	***H**yperSpectral **I**mage **C**hange **D**etection*
RNN	***R**esidual **N**eural **N**etwork*
FCN	***F**ully **C**onvolutional **N**etwork*
TLS	***T**otal **L**east **S**quares*
BP	***B**elief **P**ropagation*
AT	***A**erial **T**riangulation*
DTM	***D**igital **T**errain **M**odel*
SGM	***S**emi-**G**lobal **M**atching*
TIN	***I**rregular **T**riangulation*
SIFT	***S**cale-**I**nvariant **F**eature **T**ransform*
SURF	***S**peeded **U**p **R**obust **F**eatures*
GPU	***G**raphics **P**rocessing **U**nit*
CE	***C**ross **E**ntropy*
GIS	***G**eographic **I**nformation **S**ystem*

Chapter 1

Introduction

Change detection is the process of identifying differentiations in the state of an objector phenomenon by observing multi-temporally. Essentially, it involves the ability to quantify temporal effects using data sets acquired at divergent time point. One of the major applications of remotely-sensed data obtained from Earth-orbiting satellites is change detection because of repetitive coverage at short intervals and consistent image quality (Singh, 1989). The process uses multitemporal datasets to qualitatively analyze the temporal effects of objects or phenomena and quantify the changes (Hussain et al., 2013). Change detection studies with remote sensing data have been used in a wide variety of applications including land use and land cover, deforestation studies, natural disaster monitoring, as well as building damage assessments (Lu et al., 2004). The Due to the wide range of application scenarios including video surveillance, remote sensing, medical diagnosis and treatment, civil infrastructure, under- water sensing, driver assistance systems and so on (Roysam, 2005), imagery-based change detection related research and algorithm development has remained to be an active research focus at remote sensing and computer vision domain in recent years.

Along with the increasing diversification of applications and data categories, the technique of change detection has advanced with ascending in the spatial resolution of remote sensing images and it has been used for applications at different spatial scales. Within the applications by applying multi-temporal remote sensing imagery to generate timely information on the earth’s natural environment and human activities, most of scholars lay emphasis on natural environment related

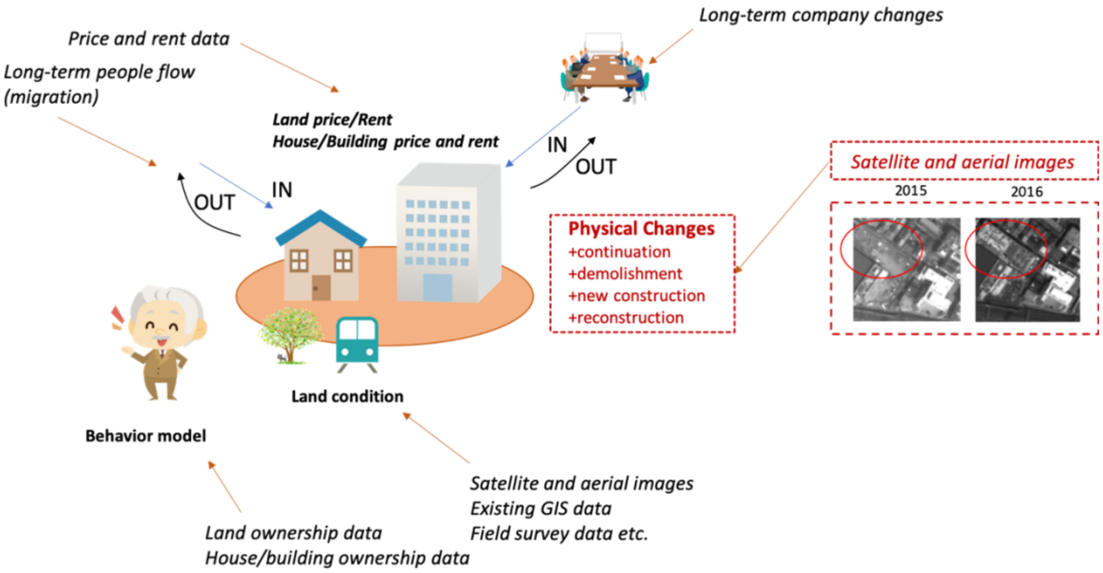


Figure 1 Research Background and Motivation of Building Change Detection

ones including monitoring of shifting cultivation, assessment of deforestation, study of changes in vegetation phenology, seasonal changes in pasture production, damage assessment, crop stress detection and so on (Singh, 1989). Nevertheless, the multi-temporal change detection of urban constructed environment, including building construction, traffic construction, urban facilities and other infrastructures, is significant for urban activities monitoring, real estate market mastery, resident's mobility supervision and then whole city development promotion. Our study will be focusing on the application scenarios of urban construction change detection, consisting of three kinds: building new construction, demolishment as well as continuation, which is aiming at urban construction legitimacy supervision and real estate commercial activity monitoring in a long term. The research background and motivation of our study is addressed in Figure 1.

Along with the developing progression of remote sensing and photogrammetry, land cover change detection is not limited to the dataset deployment on low- and medium-resolution remote sensing images based on single spectral features. Ortho urban remote sensing images derived from high-resolution aerial imagery could accommodate adequate spectral features and detailed information for feature fusion and high-level feature extraction. Our study will execute experiments on the aerial ortho imagery dataset taken from part region Tokyo, Japan in 2015 and 2016 respectively, with the resolution of 0.16m. As traditional metropolitan area, Tokyo has typical urban texture, divergent types of urban objects and high building density, which will raise the difficulty and provide solid validation for this study simultaneously.

As a premise of change detection from radiance changes from spectral feature on imagery, disturbing factors from multi-temporal aerial imagery including misalignment of pixel, radiance error caused by illumination and atmosphere condition difference should be eliminated. Although current photogrammetry methodologies could be applicated to 2-D remote sensing imagery for rectification with sensor parameters, multi-temporal aerial or satellite imagery is not adequate to offer spectral and textual features for building change detection. Alongside recent development of Dense Image Matching (DIM) technology, the acquisition of 3-D point cloud and Digital Surface Model (DSM) has been generally realized, which could be combined with imagery, making building change detection more effective with greater spatial structure and texture information.

Building change detection is a research domain, under both computer vision and remote sensing, which has developed for decades. As acknowledged in existing relevant studies, the procedure of change detection could be divided to the following steps, which is also illustrated in figure 2:

1. **Data Acquisition:** Differentiated by sensors and facilities, the data utilized for land cover change detection mainly comprises Visible Imagery (example: Aerial Imagery), Infrared Imagery as well as Multispectral, Hyperspectral and SAR Satellite Imagery, in which the last category is commonly utilized in practice.

2. **Image Pre-processing:** Likewise, depending on the sensor and data category, different image pre-processing techniques are applicated for better performance and higher efficiency in following change detection algorithm. Geometric registration, denoising and radiometric correction are basic processing methods for multi-temporal imagery data, avoiding inference factors like mis-alignment, image noise and radiance error caused by diverse perspective and illumination conditions when remote sensing equipment are working.

3. **Change Detection Algorithm:** Basically, the change detection algorithm could be categorized as pixel-based methods, object-based methods and spatial data mining methods, executed by traditional methodologies and learning-based methodologies. Relevant introduction will be given as followed and related works will be given in Section 2.

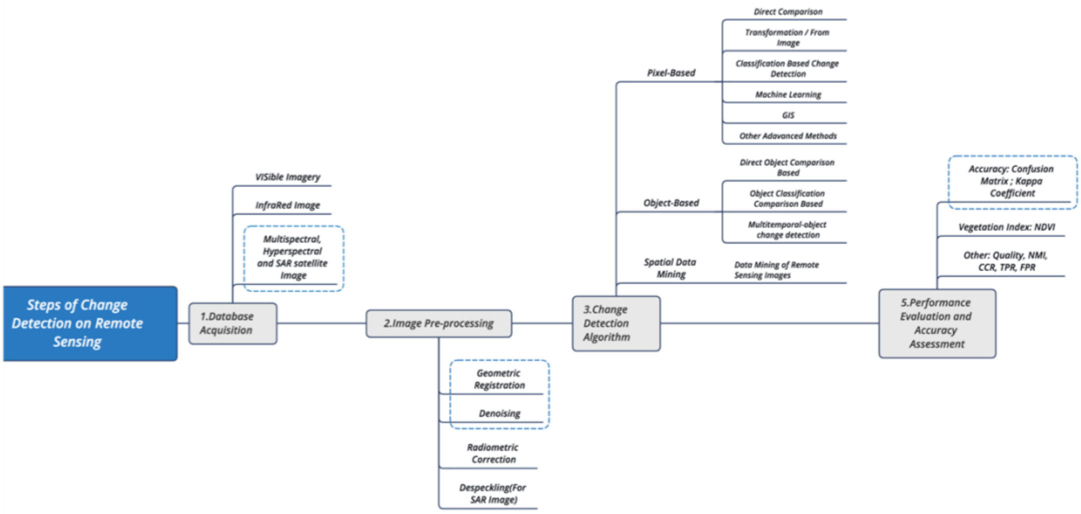

Steps of Change Detection on Remote Sensing
1.Database Acquisition
VISible Imagery
InfraRed Image
Multispectral, Hyperspectral and SAR satellite Image
2.Image Pre-processing
Geometric Registration
Denoising
Radiometric Correction
Despeckling(For SAR Image)
3.Change Detection Algorithm
Pixel-Based
Direct Comparison
Transformation / From Image
Classification Based Change Detection
Machine Learning
GIS
Other Adavanced Methods
Object-Based
Direct Object Comparison Based
Object Classification Comparison Based
Multitemporal-object change detection
Spatial Data Mining
Data Mining of Remote Sensing Images
5.Performance Evaluation and Accuracy Assessment
Accuracy: Confusion Matrix ; Kappa Coefficient
Vegetation Index: NDVI
Other: Quality, NMI, CCR, TPR, FPR

Figure 2 Framework of Change Detection Research Procedure by Current Scholars

4. **Performance Evaluation and Accuracy Assessment:** As for learning-based or model-based methods, traditional evaluation metrics including accuracy, confusion matrix, kappa coefficient and other indexes could be used for change detection algorithm performance evaluation. Also, some specialized indexes for a certain category is also frequently used. Typically, Normalized Difference Vegetation Index (NDVI) is a measure of the state of plant health based on how the plant reflects light at certain frequencies, which is used when evaluating models taking vegetation-covered area as dataset.

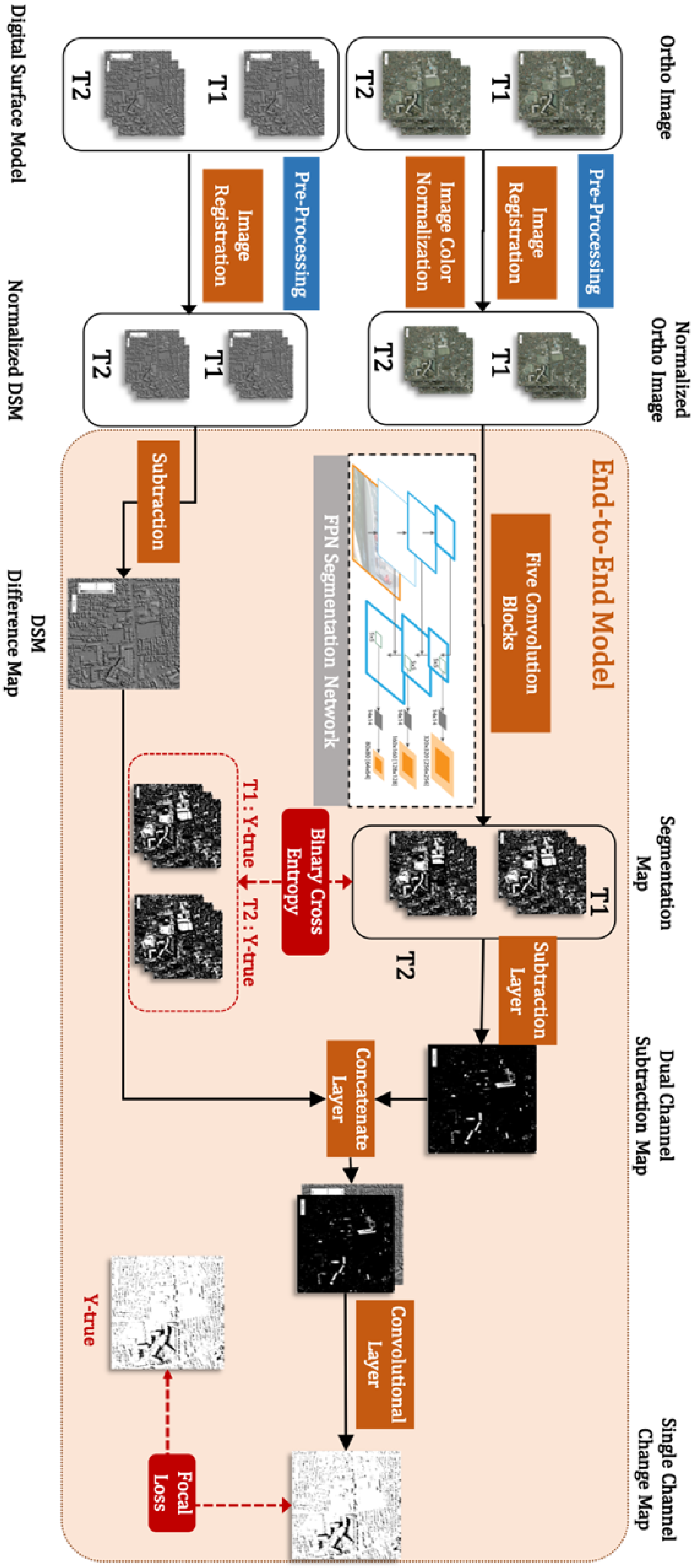


Figure 3 Overview Framework of Proposed End-to-End Change Detection Model

Current 3-D remote sensing-based change detection methods typically appertain to one of following approaches: direct comparison, classification, object-oriented method, model method and time-series analysis or hybrid method combining two or more of them (Roysam, 2005). Over the past decades, a great number of researchers and scholars have been focused on DSM-assisted change detection methods, which could be generally divided into the following three categories: (1) The first is object-based comparison, which is mostly utilized for classification and commonly employed for map updating. (2) The second is feature-based methods, in which building features including sizes and textures derived from height information are commonly utilized. (3) The third approach is to provide auxiliary change information. Although the techniques with the utilization of DSM for change detection are increasingly developed, few of them have paid attention to end-to-end model-based methods with original input of imagery and DSM scalar as the feature extraction and data utilization is challenging, due to the following factors:

1. Due to various means of 3D data generation, the uncertainty of the geometric (e.g. height) information varies with the sensors, algorithms and object scales. Uncertainties of point clouds generated using different dense matching methods may have different and non-uniform distributions. (Qin, 2016)

2. Geometric data presents a different modality from the image data. Fusion of both data requires additional consideration of data uncertainties among different types, feature extraction and multi-source weighting. The data pre-processing, data integration as well as data transformation with aerial ortho imagery is frequently limited by existing data situation, makes it a challenging data processing task.

3. High-resolution images combined with corresponding DSM requires change detection methods with the capacity of eliminating interference changes in pixels or heights, when the real land cover changes are mixed with irrelevant changes.

4. Due to the severe sample quantity difference between changed area and unchanged area, as a classification task, model-based methods are faced with sample imbalance problem. Novel loss function and appropriate adjustment on network architecture should be carried out in terms of this situation.

In view of corresponding advantages and shortcomings of great divergence of existing methods, considering the requirements of building change detection research and practical application scenario, as well as all challenging factors, our study proposes a novel end-to-end change detection model based on neural network for urban ortho aerial imagery. Instead of applying models for only classification or segmentation, the core feature extraction mechanism of end-to-end network is also appropriate for this task, mainly due to the following advantages:

1. By reducing manual preprocessing and postprocessing, end-to-end model will make the input and output to the original value, give the model more auto and independent adjustment space and increase the overall compatibility.

2. End-to-end model raises the efficiency in the elaborate problems including change detection problems where both the encoder and decoder are both trained simultaneously.

3. From the perspective of real application scenarios, the optimization code is easier to compose, and the codebase could be maintained more conveniently, because of the single framework.

With the given background and motivation above, the main objective of this study is to generate change map classified into three classes including new construction, demolishment and continuation by end-to-end model based on Feature Pyramid Network (FPN) with ortho aerial urban imagery and digital surface models (DSM) of the same target area. Specifically, the overall framework of our proposed end-to-end, multi-input and multi-output change detection model is shown in figure 3.

The rest of the paper is structured as follows: Section 2 bring forward related works for related tasks with previous informative methods. Section 3 describes the main body of proposed end-to-end change detection model architecture and framework. Model hyperparameters and specific methodology are presented in Section 4. Empirical results are discussed in Section 5, followed by conclusions and potential topics for future research in Section 6. Bibliographies and related

Chapter 2

Related Works

2.1 Related Works of Change Detection

Over the past years, scholars have put forward large numbers of change detection techniques of remote sensing image and summarized or classified them from different viewpoints. Gong has assorted change detection algorithms, a complicated and integrated process, into comparison, classification, object-oriented method, model method, time-series analysis and Hybrid methods, indicated the existing challenge over the exterior and interior steps (Gong, 2008).

2.2 Research Based on Traditional Method

As a hybrid method combining comparison and classification, Wang proposed a method based on levene-test and fuzzy evaluation especially for high-resolution remote sensing imagery, which could decrease omissions and deficiencies, improve the precision of change detection (Wang, 2018). Tian raised a method with the applying of DS fusion theory to fully use all of the change information contained in original panchromatic images, multispectral images, and the height information, for extracting real building changes. The procedure works without any learning methods or models, acquiring high calculation and running speed by traditional image-based methods (Tian, 2014). Combining pixel-based method and introducing object-based theory simultaneously, this paper acquires high rationality in every method part but low practicality because of high requirements on data quality. Xiao proposed a framework for change detection of built-up land using a combination of pixel-based change detection and object-based recognition. Specifically, the framework was derived in the sequence of detection of changed pixels by differencing and thresholding, generation of changed objects by morphological filtering, and recognition of changed objects on built-up land by object MBI difference (Xiao, 2016). Gong presented a novel change detection framework for high-resolution multispectral images, which incorporates superpixel-based change features extraction and difference representation learning model by neural networks, in which each superpixel is taken as the basic unit for change analysis for exploiting corresponding spectral, textual, and spatial features (Gong, 2017).

However, in these traditional method, inadequate data quality and fine-tuning of threshold as well as relative parameters remain to be the intrinsic challenge and drawback. For higher accuracy and lower fine-tuning manual consumption, alongside with the high-speed development of machine learning methodologies, machine learning as well as model-based deep learning methods are carried out for change detection problems gradually.

2.2.1 Research Based on Learning Method

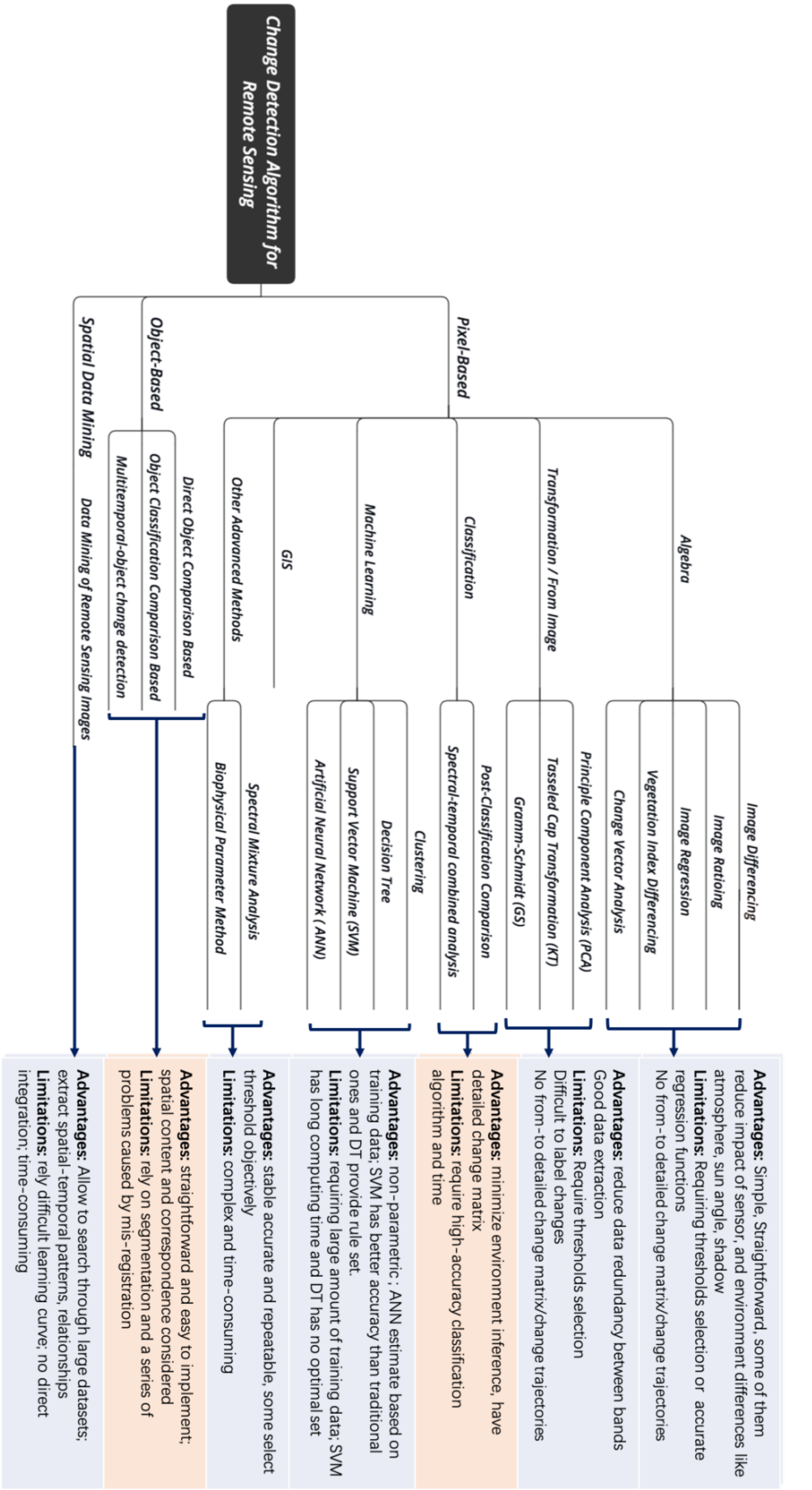
With the introduction of techniques in the domain of data science, machine learning and deep learning, neural network-based model methods, which reduces manual fine-tuning, have become a hot research direction in the past few years. Kevin utilized CNN based U-Net for semantic segmentation to extract compressed image features, as well as to classify the detected changes into the correct semantic classes, with which a difference map indicating building change information is generated as result (Kevin, 2019). The proposal of unsupervised method using pretrained model eliminates costly training process and acquires high accuracy as well as robustness. However, the separation of processing and pretrained model parameters also lead to unoptimizable model because of low comparability of corresponding tasks and datasets. Zhang presented a novel approach for change detection from multitemporal high-resolution remote sensing images without any prior information. The multitemporal deep feature collaborative learning based on SDAE (Stacked Denoising Autoencoders) is developed to obtain the deep feature representations of multitemporal images. Combined with object-level abstract difference features, a SCV (semi-supervised Chan–Vese) model will be used to extract changed regions integrated with the labeled patterns derived from an uncertainty analysis. The strength of this semi-supervised learning method mainly lays in that it could exploit the abstract difference features and improve the separability between changed and unchanged patterns. However, binary classification with uncertainty confidence could only give reference for following research but not put into real application scenarios directly (Zhang, 2019).

Deep neural networks have recently been shown to be very successful in a variety of computer vision and remote sensing tasks, which could also provide an opportunity for change detection, in which we would like to extract joint spectral-temporal features from a multi-temporal image sequence. Unsupervised methods as well as semi-supervised methods for remote sensing-based change detection could extract superficial and deep features at same time, without large preparation consumption of training data. However, the separation of segmentation process and comparison process will cause the feature and character losing. Also, the accuracy could not be guaranteed without supervision in this task, which is not clustering task or has any previous assumption as well.

2.2.2 Research Based on End-to-end Method

To eliminate the problems including inadequate model comparability, process separation and different optimizing space of pretrained weights, instead of employing model technique as one of the compositions, an end-to-end model with sufficient training dataset could be an optimal resolution. Wang presents a general end-to-end 2-D convolutional neural network (CNN) framework with the name short as GETNET for hyperspectral image change detection (HSI-CD). Mixed-affinity matrices from abundance maps obtained by linear and nonlinear spectral unmixing interacting with the HSI are processed by the GETNET. Change map as final output will be generated after another feature extraction network (Wang, 2019). This method has relatively satisfactory performance on the test dataset of natural land-cover imagery with 242 spectral bands. Although the novelty of mixed-affinity matrices provides informative data fusion technique, the unmixing technique and network architecture is not robust for other imagery data type. Mou proposed a novel neural network architecture, called ReCNN, which integrates CNN (Convolutional Neural Network) and RNN (Residual Neural Network). The network acquires the capability of extracting joint spectral-spatial-temporal features from bi-temporal multispectral images and predicts change types, which is end-to-end trainable (Mou, 2019). Nonetheless, this proposed ReCNN focused on multi-class classification task, clarifying the changes from multi-temporal aerial images to water exchange, soil exchange and city expansion, which could not meet the object-oriented demand to detect building construction change. Peng raised an improved UNet++ architecture was proposed for end-to-end CD of VHR satellite images. Dense skip connections within the UNet++ architecture was utilized to learn multi-scale feature maps from different semantic levels, with a residual block strategy to facilitate gradient convergence of the deep FCN network and capture more detailed information (Peng, 2019). Relatively good performance was shown in experiments part; however, only binary classification was implemented, which is not adequate for practical application and the network architecture was too deep and easy to encounter model degeneracy or vanishing gradients.

Comprehensive change detection model-based researches and methods were innovated, though, limited to case studies and need to be further explored for higher robustness and stronger universality. End-to-end model for change detection still has enormous development space with divergent 3D datasets, because of its great potential for feature extraction and fusion by receiving original input and generating target output without information loss. In contrast to most of the existing deep learning algorithms whose components are separately trained and tuned, it could naturally handle sequences in arbitrary lengths, referring no character segmentation or horizontal scale normalization, aiming at sequence feature extraction, higher accuracy as well as higher transfer learning flexibility. Regarding the development path of urban environment change detection task, taking the advantages, disadvantages, opportunities and challenges of the multi-source data and current techniques, we are aiming at proposing a novel end-to-end model-based method.



Change Detection Algorithm for Remote Sensing

Pixel-Based

Algebra: Image Differencing; Image Rationing; Image Regression; Vegetation Index Differencing; Change Vector Analysis

Advantages: Simple, Straightforward, some of them reduce impact of sensor, and environment differences like atmosphere, sun angle, shadow
Limitations: Requiring thresholds selection or accurate regression functions
No from-to detailed change matrix/change trajectories

Transformation / From Image: Principle Component Analysis (PCA); Tasseled Cap Transformation (KT); Gramm-Schmidt (GS)

Advantages: reduce data redundancy between bands
Good data extraction
Limitations: Require thresholds selection
Difficult to label changes
No from-to detailed change matrix/change trajectories

Classification: Post-Classification Comparison; Spectral-temporal combined analysis

Advantages: minimize environment inference, have detailed change matrix
Limitations: require high-accuracy classification algorithm and time

Machine Learning: Clustering; Decision Tree; Support Vector Machine (SVM); Artificial Neural Network (ANN)

Advantages: non-parametric ; ANN estimate based on training data; SVM has better accuracy than traditional ones and DT provide rule set.
Limitations: requiring large amount of training data; SVM has long computing time and DT has no optimal set

GIS

Other Adavanced Methods: Spectral Mixture Analysis; Biophysical Parameter Method

Advantages: stable accurate and repeatable, some select threshold objectively
Limitations: complex and time-consuming

Object-Based: Direct Object Comparison Based; Object Classification Comparison Based; Multitemporal-object change detection

Advantages: straightforward and easy to implement; spatial content and correspondence considered
Limitations: rely on segmentation and a series of problems caused by mis-registration

Spatial Data Mining: Data Mining of Remote Sensing Images

Advantages: Allow to search through large datasets; extract spatial-temporal patterns, relationships
Limitations: rely difficult learning curve; no direct integration; time-consuming

Figure 4 Framework of Current Existing Methodologies of Change Detection Based on Remote Sensing

2.3 Related Works of Semantic Segmentation

As a multi-task model-based method, our proposed method is an approach realizing multi-temporal urban change detection and building semantic segmentation simultaneously. In several remote sensing imagery and DSM based building segmentation algorithms, the 3D points are classified in three classes: terrain, clutter and building. First, the 3D measurements are classified as ground and non-ground, and subsequently the non-ground points are divided into clutter and building regions (Matei, 2008).

Kraus and Pfeifer presented an iterative method for terrain modeling based on removing at each step the 3D measurements with residuals to the current terrain surface larger than a threshold and re-estimating the terrain surface using the remaining data (Kraus, 1998). Because the initialization of the terrain used all the data, the method may not converge for densely built regions. Considering the significant existence of urban area of high-density constructions and population, like Tokyo and Shanghai, the boundedness of this method may cause the difficulty in being taken into practice. Morphological opening operators are used to create a digital terrain model (DTM) which is subtracted from the input data. These filters, inspired from image processing may fail to produce good ground segmentation, especially for nonfiat terrain (Rottensteiner, 2002). Verma segmented the ground and the buildings in one step by computing local planar patches using Total Least Squares (TLS) and connecting the consistent neighboring surfels into regions using bottom-up region growing (Verma, 2006). The largest region was selected as ground, while the rest of the regions were classified as individual buildings.

Rottensteiner et al. presented a method for the automatic delineation of roof planes using local plane estimation and grouping the local planes into larger regions starting from local seed locations. Over-segmented regions are merged together using co-planarity constraints. Points that do not belong to any surface are labeled as clutter. Belief propagation (BP) is applied for segmentation tasks with a similar formulation as ours. For example, Sun et al. used a max-product BP algorithm for stereo matching to enforce constraints that neighboring pixels with the same intensity values are assigned the same depth (Sun, 2003). Guo et al. employed a rectilinear approximation to outlines of buildings in their image-based building detection system (Guo, 2001). The orientation of each rectilinear fit was determined by the maximum in local image gradients for tangent directions to the contour.

Traditional methods and learning-based methods for building semantic segmentation have developed for several decades and become an advanced technique, which has already been taken into diversified urban application scenarios. In our research, which is a multi-task model-based method, building segmentation map will be generated as additional output, assisting the change detection and avoiding the great bias caused by sample imbalance problem, whose accuracy is also greatly essential.

Chapter 3

End-to-end Change Detection Model

The overview framework of our proposed end-to-end change detection model is shown in figure 5. Generally, the proposed framework comprises three components:

- DSM generating with raw aerial image by photogrammetry algorithm (traditional algorithms)
- Image pre-processing including image registration, radiometric balance and image color normalization (traditional algorithms)
- End-to-end dual change detection model based on Feature Pyramid Network (FPN). (Neural Network algorithms)

The model architecture and composition principle will be given in this section as followed and the pre-processing techniques and algorithm explanation as well as the detailed implementation of all three components will be given in the next Section of Methodology, including pre-processing and post-processing methods, network architecture and advantages, as well as our proposed novel focal loss.

As the major dual change detection deep learning model, based on CNN (Convolutional Neural Network), we adapt FPN (Feature Pyramid Network) to implement domain feature extraction on multi scales for urban objects, mainly for urban constructions. The model is modified to receive four inputs including pre-processed ortho images and pre-processed DSMs of T1 (Time1) and T2 (Time2), containing the same urban area, and then generate three outputs including building segmentation maps of T1 and T2 and the multi-class change map containing the changed objects from T1 to T2, which is our main output and main research objective. In this model, simultaneously, a binary-classification building semantic segmentation problem and a multi-classification building change detection problem are trained together in an end-to-end model.

As the training dataset generator, normalized ortho imagery and DSM will be clipped into patches with the width and height of $(224, 224)$. The RGB imagery has 3 channels will be transformed to tensor in $(N_{patch}, 224, 224, 3)$ with value in $[0, 255]$ and the DSM will become tensor in $(N_{patch}, 224, 224, 1)$ with only 1 channel as elevation value in $[-1, 1]$, where N_{patch} represents the number of patches. The ground truth dataset, including building segmentation map and change map share the same original size of width and height, as well as patch number.

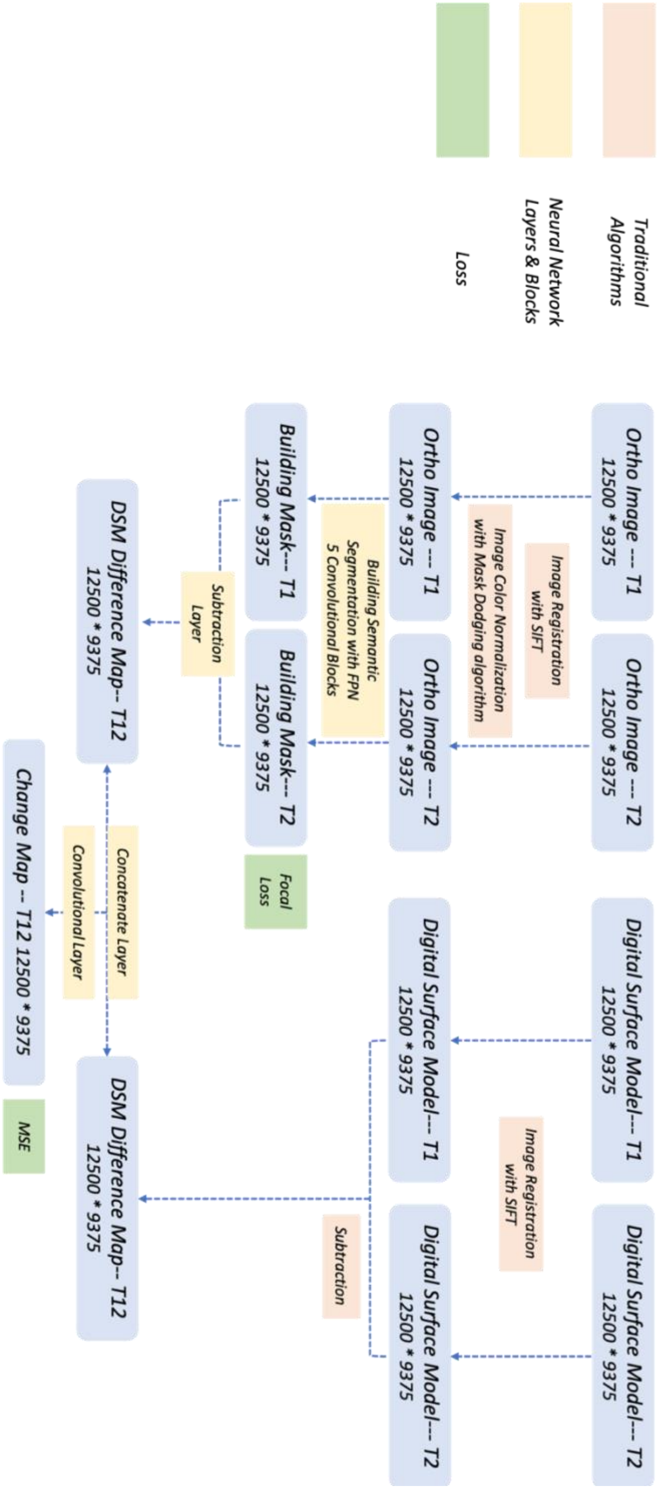

Traditional Algorithms
Neural Network Layers & Blocks
Loss
Ortho Image --- T1 12500 * 9375
Ortho Image --- T2 12500 * 9375
Image Registration with SIFT
Image Color Normalization with Mask Dodging algorithm
Ortho Image --- T1 12500 * 9375
Ortho Image --- T2 12500 * 9375
Building Semantic Segmentation with FPN 5 Convolutional Blocks
Building Mask--- T1 12500 * 9375
Building Mask--- T2 12500 * 9375
Focal Loss
Subtraction Layer
DSM Difference Map-- T12 12500 * 9375
Concatenate Layer
Convolutional Layer
Change Map -- T12 12500 * 9375
MSE
Digital Surface Model--- T1 12500 * 9375
Digital Surface Model--- T2 12500 * 9375
Image Registration with SIFT
Digital Surface Model--- T1 12500 * 9375
Digital Surface Model--- T2 12500 * 9375
Subtraction
DSM Difference Map-- T12 12500 * 9375

Figure 5 Overall Method Decomposition of Proposed Method

The framework of shape transformation processing and feature composition is shown in figure 6. As the segmentation is binary classification problem, the generator will make them tensors in the shape of $(N_{patch}, 224, 224, 2)$ with one-hot encoding and the change map ground truth will be transformed to $(N_{patch}, 224, 224, 3)$ with 3 classes. In the process of DSM, a subtraction layer operates element-wised subtraction between input DSM T1 tensor and DSM T2 tensor and generate a tensor also with the shape of $(N_{patch}, 224, 224, 1)$ in $[-1, 1]$. The intermediate output from FPN will be two binary classified tensors with shape $(N_{patch}, 224, 224, 2)$ and value in $\{0,1\}$. Later a subtraction layer merges these two tensors with element-wised subtraction and get one tensor with $(N_{patch}, 224, 224, 2)$ shape and value in $\{-1, 0,1\}$. A concatenate layer will concatenate DSM subtracted tensor and segmentation subtracted tensor in $axis = 3$ then generate tensor in $(N_{patch}, 224, 224, 3)$ combined imagery low-level feature, high-level feature and DSM feature. It will go through a convolutional block comprises 3 convolutional layers with (3,3) kernel, 32 channels and same padding, then output change map tensor in $(N_{patch}, 224, 224, 3)$ as final output. The spatial size as well as resolution will remain same with original inputs.

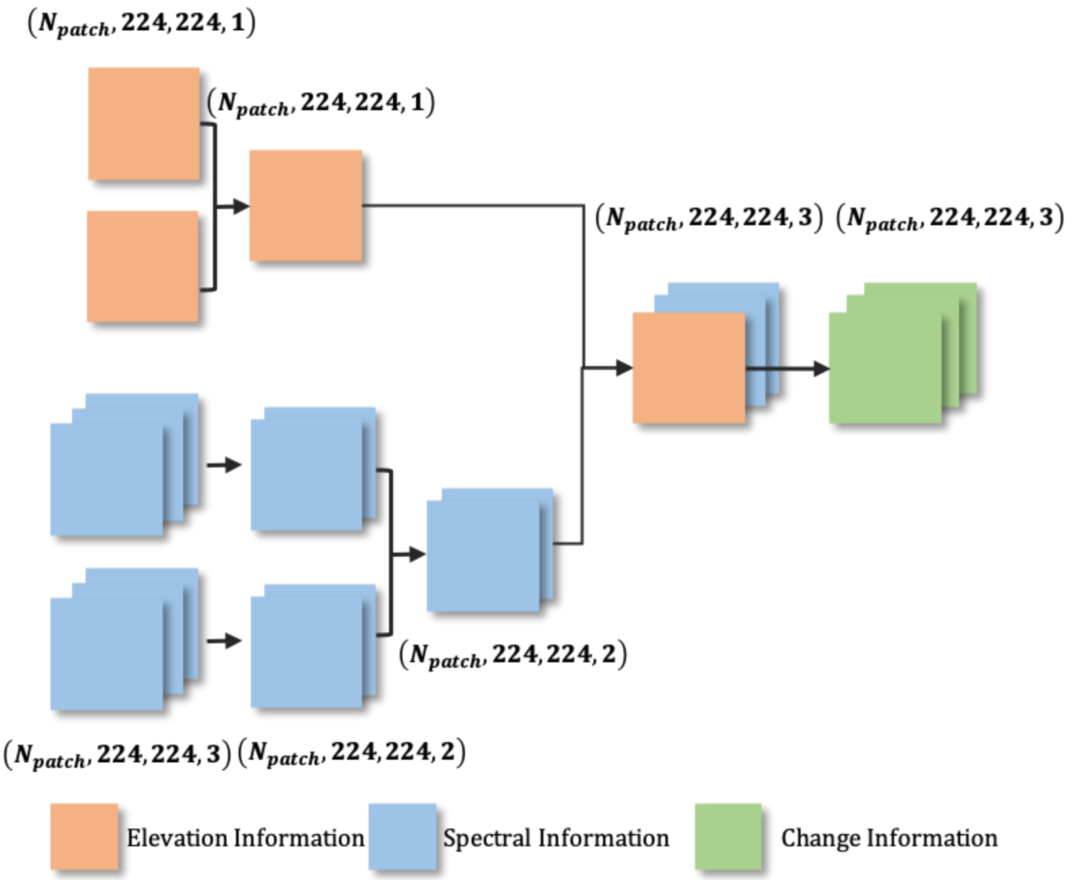


Figure 6 Shape Transformation and Feature Composition of Proposed Neural Network

The loss function is another crucial part except for model architecture, which comprises binary cross entropy loss for two segmentation intermediate outputs and focal loss for final output as change map, with equal loss weights setting. The combination of diversified losses could guarantee the building modality extraction and learn the change detection pattern at the same time, without miss-classifying tendency of model or low predictive performance towards the minority class.

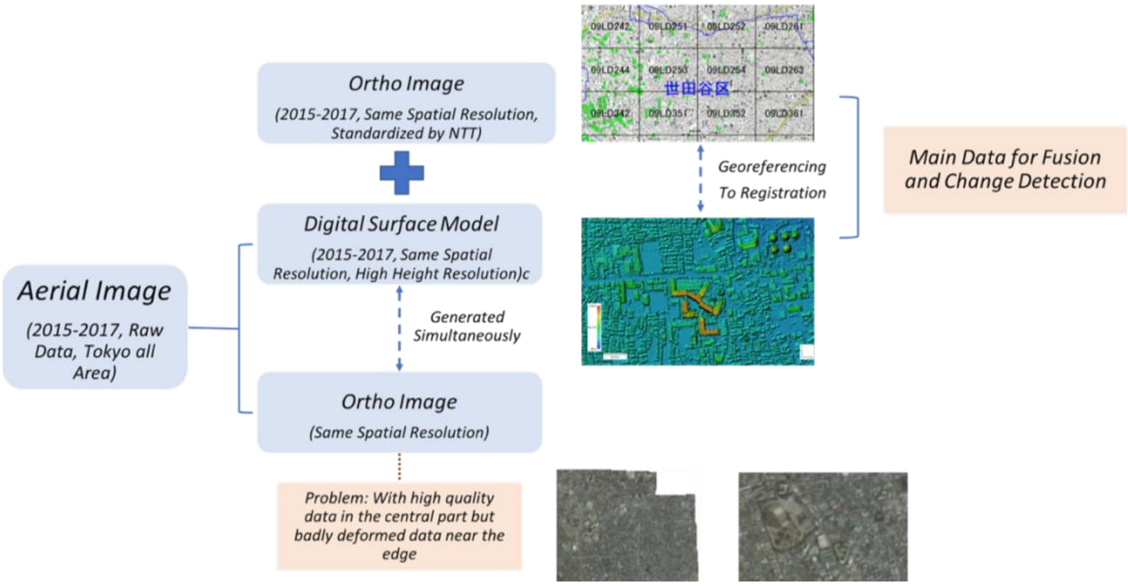
The biggest challenge in the training process in this model is the severe label imbalance in change map prediction, with the ratio between the amount of continuation label and the new construction/demolishment label over 1000. Under existing circumstances, balancing approaches from the perspective of both data and model should be brought out to avoid the imbalance problem biasing the model, causing miss-classifying towards majority class and low predictive performance on minority class particularly. In this way, the significance of cross entropy losses is to optimize the segmentation, aiming at providing building foreground and background feature to keep the object semantics but not just final pixel-wised prediction. Also, the focal loss is an efficient loss function targeting at training data label imbalance problem and object-oriented problem, which will be introduced in next Section particularly about its mathematical basis and our modifications.

Chapter 4

Methodology

This section discusses the methodology utilized to conduct this research. In Section 4.1, photogrammetry related techniques are introduced, which were conducted for the transformation of aerial imagery raw data to ortho image and DSM. Section 4.2 describes the image-preprocessing measures including image registration and radiometric correction employed in this study. Section 4.3 discusses the characters of FPN as feature extraction network and its superiority for change detection task. Section 4.4 describes the calculation and theory of the novel multi-class focal loss as a crucial loss function used in model training optimizer.

4.1 Photogrammetry



09LD242
09LD251
09LD252
09LD261
09LD244
09LD253
09LD254
09LD263
世田谷区
09LD242
09LD351
09LD352
09LD361
Ortho Image
(2015-2017, Same Spatial Resolution, Standardized by NTT)
Georeferencing
To Registration
Main Data for Fusion and Change Detection
Digital Surface Model
(2015-2017, Same Spatial Resolution, High Height Resolution)c
Aerial Image
(2015-2017, Raw Data, Tokyo all Area)
Generated Simultaneously
Ortho Image
(Same Spatial Resolution)
Problem: With high quality data in the central part but badly deformed data near the edge

Figure 7 Data Composition and Acquisition Procedure

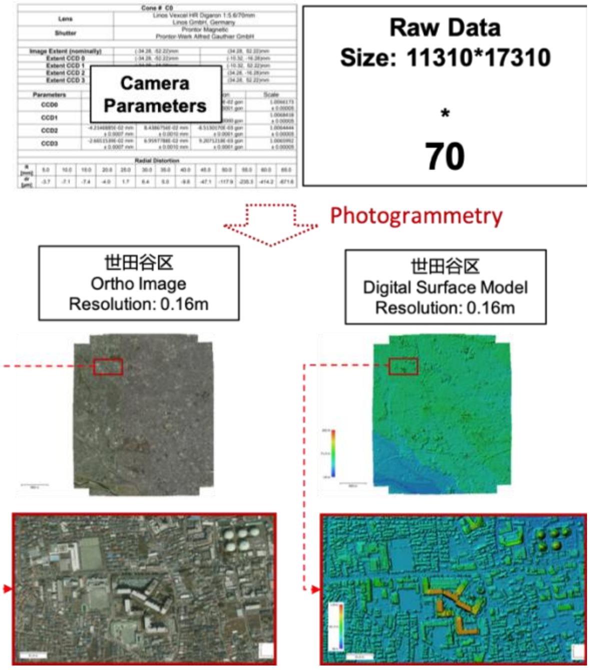


Figure 8 General process of DSM and Ortho image Acquisition from Raw Data

Figure 7 shows the general process of DSM and ortho image acquisition from raw data in our research, in which Photogrammetry is a significant part and technique. The function and working flow of photogrammetry utilized in our research is addressed in figure 8.

Photogrammetry can be defined as the study domain of determining qualitative and quantitative features of objects taken from the images recorded on photographs, no matter hardcopy film or digital imagery. Identification and qualitative description of objects could be done by observing the characteristics of photographic images (such as shape, pattern, tone, and texture). The identification of tree types, the description of land cover textures, and the existing land use inventories are examples of qualitative observations obtained through photogrammetry. The characteristics of objects (such as size, direction, and position) are determined from the measured image position in the camera image plane by photogrammetry. Tree height, storage capacity, topographic map, and horizontal and vertical coordinates of unknown points are examples of quantitative measurement results obtained from photography. The following is a review of the basic geometry of aerial imagery and the photogrammetric elements that form the basis of a photogrammetry solution. Comparisons are also provided between analog, analytical and digital photogrammetry.

The aerial photogrammetry method can use the 2D coordinates measured on the stereo aerial photos (see the same point on the ground from two or more different angles) to accurately draw the position of the 3D coordinates on the ground. Figure 9 illustrates how to determine the 3D ground coordinates of point P from the 2D photo coordinates of p1 and p2 measured from a pair of stereo aerial photos by photogrammetry.

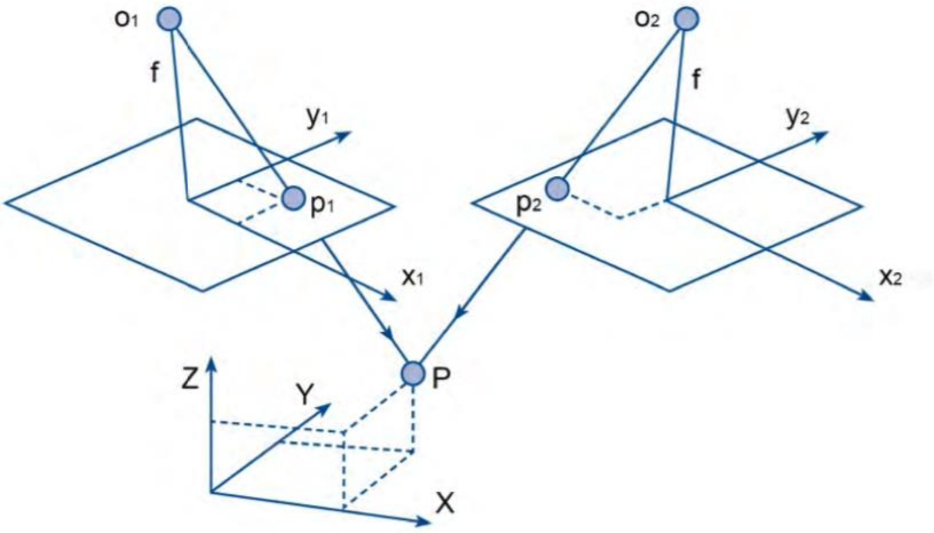


Figure 9 Principles of Photogrammetry – how 2-D measurements of points p_1 and p_1 on stereo photograph enable the 3-D mapping of point P on the ground

In figure 9, f is the focal length of the camera lens; $o_1$and o_2 represent the focal position of the camera lens when taking pictures 1 and 2; x_1 and x_2 depict the flight direction when each image is acquired and the $\mathrm{x-axis}$ is established on the stereo picture; y_1 and y_2 show the cross flight direction and establish the $\mathrm{y-axis}$ on the stereo photo. p_1 and p_2 show the location of the point p imaged on the two photos.

The photo coordinate system establishes x_1 and y_1 photo coordinates of point p_1 on the left image and establishes x_2 and y_2 photo coordinates of point p_2 on the right image. The main point to remember from the short summary of photogrammetry is that light rays project from o_1 through p_1 (inside the camera in the image on the left) to P (on the ground), and light rays project from o_2 through p_2 on p_2 (internal). Camera to get the correct image) to P (on the ground), will never intersect at P (to determine its X/Y/Z coordinates on the ground) unless there is (1) a good internal direction to define The geometric parameters of the imaging process inside the camera and (2) good external direction, this angle defines the exact 3D position of the camera lens focus in the air (x/y/z coordinates of points o_1 and o_2) and the precise angular direction (Scroll to obtain the pitch and yaw angles of the airplane and camera when acquiring each photo or digital image).

Camera calibration establishes internal orientation parameters for each lens cone of a metric camera, while aerial triangulation (AT) establishes external orientation parameters and absolute orientation of all stereo models for each photo or digital image to accommodate ground control. The construction of the metric camera keeps its image characteristics stable.

Creating a digital surface model (DSM) or digital terrain model (DTM) requires collecting topographic polylines, which are linear features that describe changes in surface smoothness or continuity. The quality point is a separate 3D point, which is also placed three-dimensionally on the ground at an equidistant position, and there is no terrain break line. Quality points are usually generated by a process called semi-global matching (SGM) based on automatic image correlation techniques. When a contour is to be generated, the position of the particles will change according to the required contour interval. The difference between DSM and DTM is that DSM stands for the top reflective surface, including vegetation, bridges and buildings as well as terrain boundaries and particles. DTM is a bare model, in which all surface objects have been deleted, leaving only the ground terrain. Normally, only terrain contour lines and quality points are collected to create a DTM. Once the area of interest has topographic polylines and quality points, the data can be extracted into the GIS software, which can generate a TIN (irregular triangulation). You can then create an outline from the TIN file and view it in stereo software. Viewing the outline on the image in a stereoscopic way allows the analyst to see where additional polylines or mass points may be needed, whether the outline floats or cuts into the hillside. With modern soft copy photogrammetry, it is no longer necessary to manually edit contour lines by moving floating points along the ground at a set height.

For the generation of ortho imagery, orthophotos are generated using special software, corrected (pixel by pixel), and spatially located on the surface of the earth. The orthophoto software uses camera calibration files, AT solutions, digital elevation files and original photos to generate orthophotos. After creating a single orthophoto, you can mosaic them into a seamless image or multiple combined images based on a square tile grid (USGS Quad / Quarter Quad border or user-specified dimensions). The corrected images should be quality checked to ensure that they are free of voids, sensor artifacts and distorted features. Deformation characteristics are usually caused by errors in the elevation model or by bridges corrected to the bare soil model. When generating orthophotos and mosaics, the file size should be considered. In our study, due to the orthophoto image problems generated by the original aeronautical data, including the deformation and misalignment of the orthophoto image edge area, we are using the orthophoto image provided by NTT.

4.2 Image-Preprocessing

4.2.1 Image Registration

As a result of the differentiation in camera condition, illumination, perspectives and other inference factors, the positions of corresponding buildings on multi-temporal ortho images could not align with each other accurately. As shown in Figure 10, when we overlap 2015 image and 2016, it shows slight misalignment of low-rise Building area but crucial misalignment of high-rise building area.

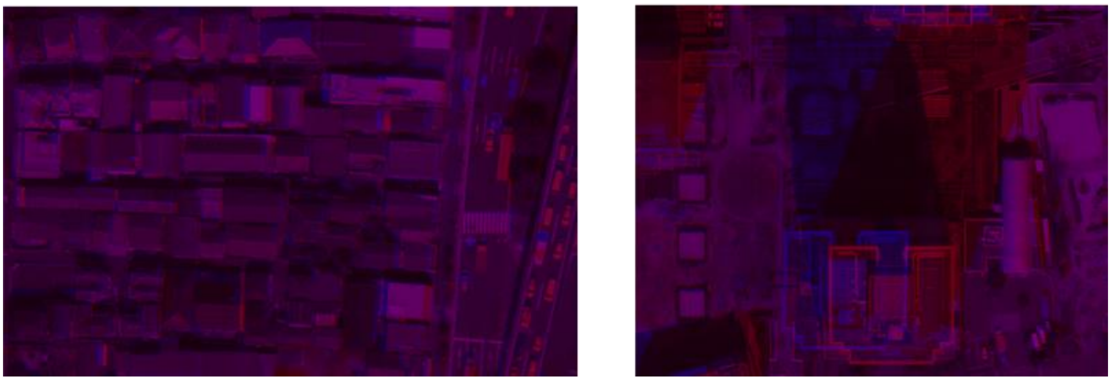


Figure 10 Overlapping Maps of 2015 ortho image (blue channel) and 2016 (red channel)

Image registration aims at integrating multi-temporal aerial ortho image into optimal geometric alignment and georeferencing condition, which is widely used in a variety of applications in remote sensing field. The framework of existing image registration algorithms we summarized is shown as figure 11, we choose Scale-Invariant Feature Transform (SIFT) raised by Low in 2004 and Speeded Up Robust Features (SURF) for experiments and SIFT shows better performance. Therefore, basically, we adapt SIFT algorithm to corresponding patches of original image for feature key points localization and matching.

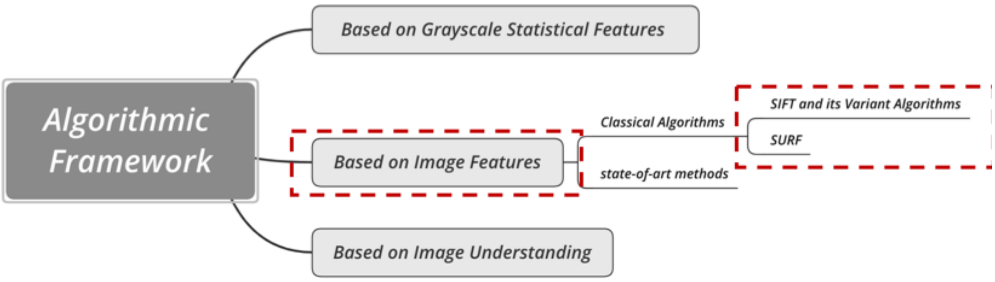

Algorithmic Framework
Based on Grayscale Statistical Features
Based on Image Features
Classical Algorithms
SIFT and its Variant Algorithms
SURF
state-of-art methods
Based on Image Understanding

Figure 11 Framework of Existing Image Registration Algorithms

SIFT algorithm is short for Scale Invariant Feature Transform, as it transforms image data into scale-invariant coordinates relative to local feature (Low, 2004). Basically, it could be divided into the following major steps of computation used to generate the set of image features:

1. **Scale-space extrema detection**: The first stage of calculation searches for all scales and image positions. By using the Gaussian difference function to identify potential points of interest with constant size and direction, this goal can be effectively achieved. The algorithm uses a cascade filtering method to detect key points. This method uses an effective algorithm to identify candidate positions and then examines them in detail. The first step of key point detection is to identify the positions and scales that can be repeatedly specified in different views of the same object (see figure 12). By using a continuous function of scales called scale space, searching for stable features on all possible scales, it is possible to detect positions where the scale of the image does not change.

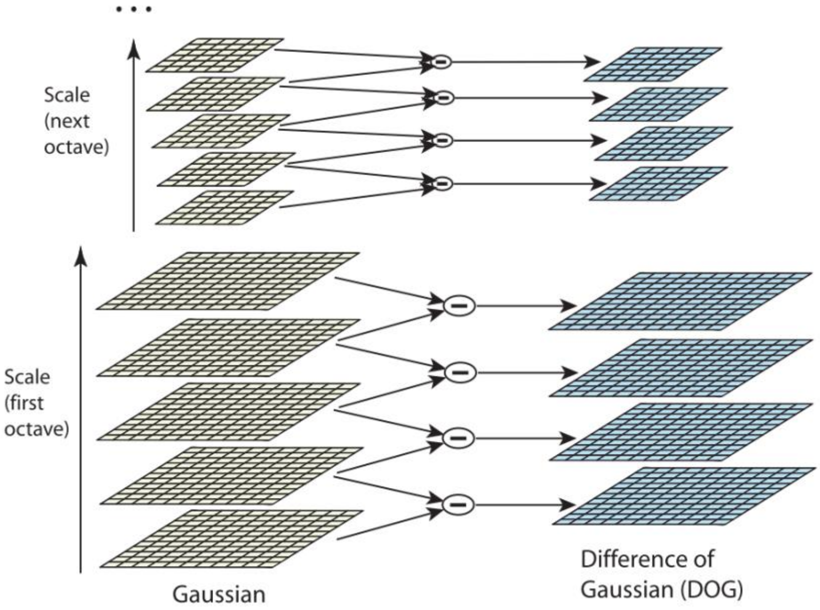


Figure 12 Generation of difference-of-Guassian images

For each octave of scale space, the initial image is repeatedly convolved with Guassians to produce the set of scale space images shown on the left. Adjacent Guassian imaged are subtracted to produce the difference-of-Guassian images on the right. After each octave, the Guassian image is down-sampled by a factor of 2, and the process repeated.

2. **Keypoint localization**: At each candidate location, it is appropriate to use a detailed model to determine the location and scale. Select key points according to their stability. In order to detect the local maximum and minimum values of D(x, y, σ), it can be calculated from the difference between two nearby scales, each sample point and its eight neighbors in the current image and nine of the scales The neighbors are compared above and below (see Figure 13). It will be selected only if it is greater than all these neighbors or less than all neighbors. Since most sample points will be eliminated after the first few inspections, the cost of this inspection is quite low.

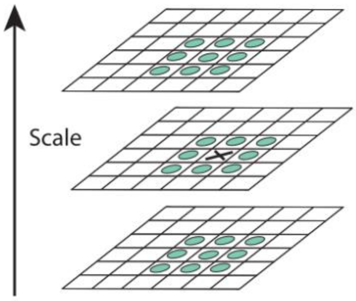


Figure 13 Maxima and minima of the difference-of-Gaussian images are detected by comparing a pixel (marked with X) to its 26 neighbors in 3x3 regions at the current and adjacent scales (marked with circles).

3. **Orientation assignment**: According to the local image gradient direction, assign one or more directions to each key point position. All future operations are performed on the image data that has been converted with respect to the direction, scale, and position assigned to each feature, thereby providing invariance for these conversions.

4. **Keypoint descriptor**: The local image gradient is measured at the selected scale in the area around each key point. These are converted into representations, allowing significant levels of local shape distortion and lighting changes. The previous operation has assigned the image position, scale and orientation to each key point. These parameters impose a repeatable local 2D coordinate system in which the local image area can be described, so these parameters can be provided with invariance. The next step is to calculate a descriptor for the local image area. The descriptor is highly distinguishable, but the remaining changes such as changes in lighting or 3D viewpoint are kept as constant as possible.

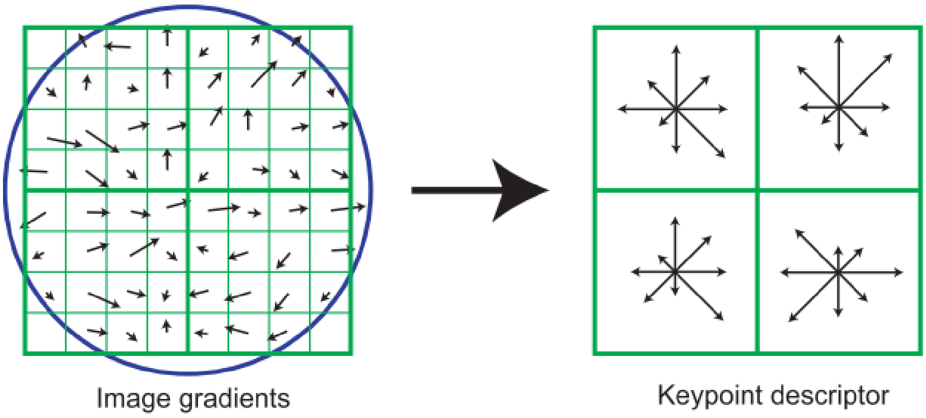


Figure 14 Keypoint Descriptor

A keypoint descriptor is created by first computing the gradient magnitude and orientation at each image sample point in a region around the keypoint location, as shown on the left. These are weighted by a Gaussian window, indicated by the overlaid circle. These samples are then accumulated into orientation histograms summarizing the contents over 4x4 sub regions, as shown on the right, with the length of each arrow corresponding to the sum of the gradient magnitudes near that direction within the region. This figure shows a 2x2 descriptor array computed from an 8x8 set of samples, whereas the experiments in this paper use 4x4 descriptors computed from a 16x16 sample array.

Figure 14 illustrates the calculation of keypoint descriptors. First, use the scale of the key points to sample the size and direction of the image gradient for the positions around the key points. In order to achieve direction invariance, the coordinates and gradient direction of the descriptor are rotated relative to the key point direction. It is worth mentioning that a Gaussian weighting function with σ equal to half the width of the descriptor window is used to assign weights to the size of each sampling point. This is indicated by a circular window on the left side of figure 14, although of course the weight will drop steadily. The purpose of this Gaussian window is to avoid sudden changes in the descriptor and small changes in the position of the window and reduce the emphasis on gradients away from the center of the descriptor, because these gradients are most susceptible to poor registration errors.

The keypoint descriptor is shown on the right side of figure 14 It allows for significant shift in gradient positions by creating orientation histograms over 4x4 sample regions. The figure shows eight directions for each orientation histogram, with the length of each arrow corresponding to the magnitude of that histogram entry. A gradient sample on the left can shift up to 4 sample positions while still contributing to the same histogram on the right, thereby achieving the objective of allowing for larger local positional shifts.

SIFT keypoints are particularly useful for our image registration problems because of their uniqueness, which makes it possible to select the correct match of keypoints from a large database of other keypoints. This uniqueness is achieved by combining high-dimensional vectors representing image gradients in local regions of the image. The key points have been shown to be unchanged for image rotation and scaling and are robust to a large range of affine distortion, noise increase, and lighting changes. A large number of key points can be extracted from a typical image, making it robust when extracting small objects from clutter.

The fact that key points could be detected over the entire scale range denotes that, smaller local features can be used to match smaller and highly occluded objects, while larger key points work well on images affected by noise and blur it is good. They are computationally efficient, so they can extract thousands of key points from typical images with near real-time performance on standard PC hardware (David, 2004). Figure 15 shows the keypoint matching results of orthogonal aerial images of the same area in Tokyo in 2015 and 2016. The green line connects the corresponding matching key points and generates a linear homography matrix for the entire image transformation for image registration.

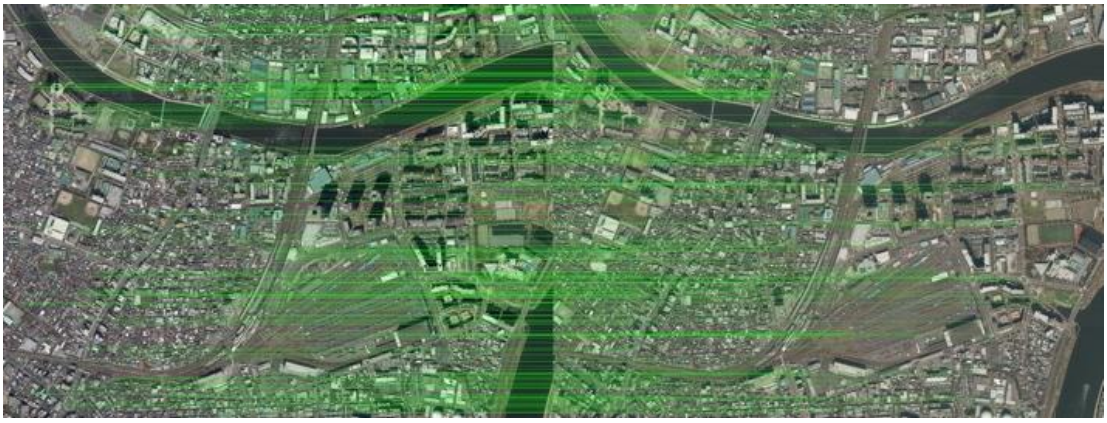


Figure 15 Matched Key points with SIFT on Tokyo Region

4.2.2 Radiometric Correction

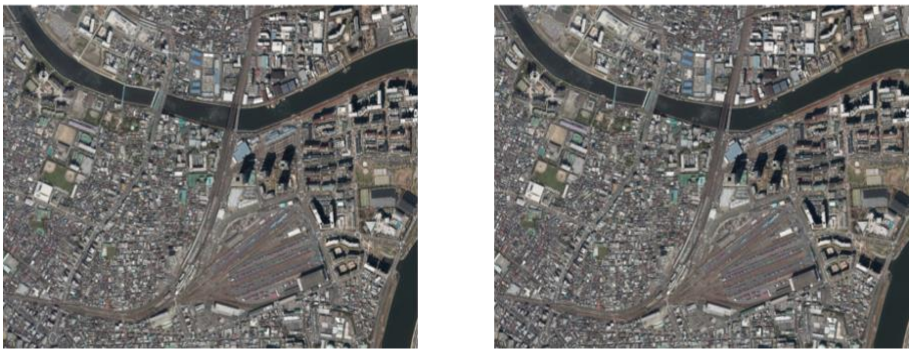


Figure 16 Radiometric Condition of 2015 and 2016 Aerial Images Comparison

In view of the fact that the layer parameters of the T1 and T2 input orthogonal images will be shared in the feature extraction network, the radiation divergence of the aerial image caused by interference factors will become one of the important factors during the training of the neural network (see figure 16). In order to unify the differences in color balance conditions caused by the imaging season or date, different sun heights and illuminations, different angles, different meteorological conditions, and different cloud, rain, or snow coverage areas, to optimize the model performance, which improves Accuracy and effectiveness, radiation correction and color normalization methods.

Radiometric correction could eliminate the impact of changing the spectral characteristics of land features, except for actual changes in ground targets, which have become mandatory measures in multi-sensor and multi-date studies (Paolini, L., 2006). Satellite image radiation correction methods can be divided into two categories: absolute values and relative values (Thome et al., 1997). In our research, the relative radiation correction method is relatively suitable for multi-time radiation divergence of aerial images.

Relative radiation correction can be used to eliminate or normalize changes between images and generate radiation-standardized data at a common (reference) scale. It is also used to normalize the differences between the various detectors and is usually done in the state of product generation (Du, 2002). The relative radiation correction method normalizes images of the same area and different dates by using landscape elements (pixels) whose reflectance values are almost constant over time. This process assumes that the pixels sampled at T2 are linearly correlated with the pixels sampled at T1 at the same location, and that the spectral reflectance characteristics of the sampled pixels do not change during the time interval. The sampled pixels are regarded as pseudo-invariant features (PIF) and are the key to the image regression method used in the normalization process. The main characteristics of PIF are that they are considered to be spatially well-defined objects and remain stable in the spectrum over time. The limitation of this method is that the landscape elements are usually selected through visual inspection, which may lead to subjective radiation standardization. (Paolini, L., 2006).

In order to perform relative radiation correction, Du proposed an objective method for selecting PIF. The main assumption of this method is that the linear effect that affects the image is much larger than the nonlinear effect, so

$$Q = La + b \quad (1)$$

where Q is the image value in digital counts, L is the surface radiance of the imaged scene, and a and b are linear coefficients that take changes in satellite sensor calibration over time, differences in illumination and observation angles, atmospheric effects and so on into consideration.

From equation (1), it can be shown that the statistical properties of the PIFs are constants for all the images. There is an attribute $A(i)$ of each PIF, independent of the image characteristics,

$$\frac{(Q(i)-\bar{Q})^2}{\frac{1}{n}\sum_{i=1}^{l=n}(Q(i)-\bar{Q})^2} = \frac{(L(i)-\bar{L})^2}{\frac{1}{n}\sum_{i=1}^{l=n}(L(i)-\bar{L})^2} = A(i) \quad (2)$$

It can be seen, from equation (2), that $A(i)$ is a factor that is independent of the a and b coefficients and dimensionless. Therefore, $A(i)$ represents a property of the PIFs that is independent of all linear variations affecting the image. By selecting PIF, you can establish a linear effect that affects the image, and therefore determine the correction factor to be applied to the standardized image. The PIF selection procedure proposed by Du is an objective process based on PCA to calculate similar band pairs of different images, involving:

1. Applying thresholds values to each band of each image to reject cloudy and water pixels.

2. Using the remaining pixels to compute the PCA between each pair of bands of each image.

3. Then, the pixels located around the primary major axis will be selected as PIFs, using an arbitrary threshold U perpendicular to the PCA major axis.

Once the PIFs are selected, the mean and standard deviation of each band in each image are calculated, and the gain and offset to normalize the images are computed as,

$$\text{gain}_{(j)} = \frac{\sigma Q_{ref}}{\sigma Q_j} \quad (3)$$

$$\text{offset}_{(j)} = \bar{Q}_{ref} - \frac{\sigma Q_{ref}}{\sigma Q_j} \cdot \bar{Q}_j \quad (4)$$

where j represents the image date, $\bar{Q}_{ref}$ and σQ_{ref} are the references mean and standard deviation values, respectively, and $\bar{Q}_i$ and σQ_i are the mean and standard deviation of each set of PIFs. In this way, by applying new gain and offset values for each frequency band of each image, the entire image set is normalized to a reference ratio common to all images.

4.3 Feature Pyramid Network (FPN)

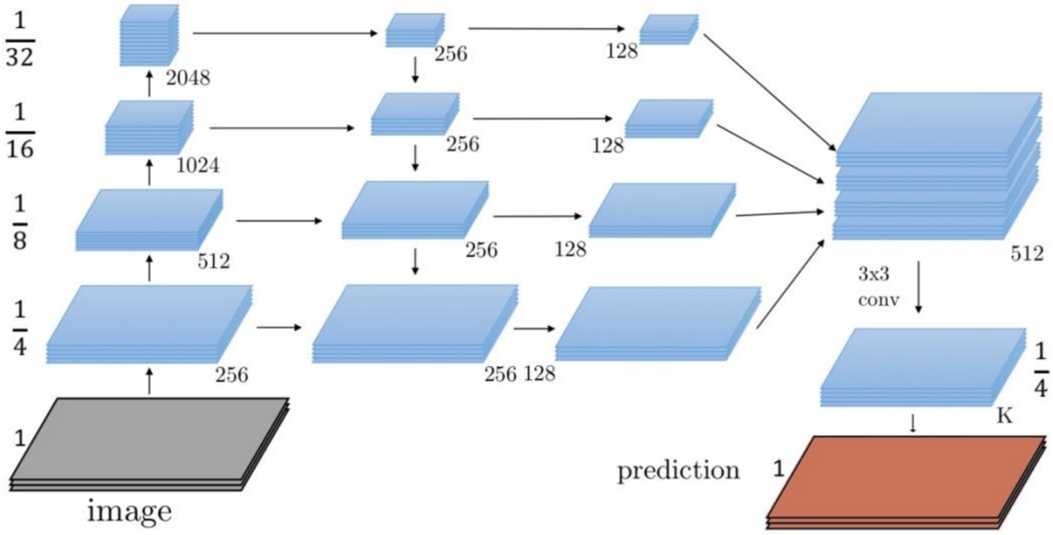


Figure 17 Architecture of FPN Component Utilized in Proposed Model

Lin published Feature Pyramid Network (FPN) in 2019, which was an outstanding academic outcome in computer vision domain, showing excellent performances on object detection task (Lin, 2019). In our research on building change detection, which is pixel-wised classification task, the network is utilized for semantic segmentation task. Keeping all the original mechanism and feature extraction method, some modification was executed on the network architecture, which is demonstrated in figure 17.

4.3.1 Deep Convolutional Networks (ConvNets)

For FPN, the basis of this model-based approach is CNN and deep learning, which includes a great variety of neural network-based models. An important part of the FPN hierarchy is the ConvNet layer, which follows the CNN calculation model.

The ability to control CNN-based models can be controlled by changing the depth and breadth of the CNN model, and they also make a strong and almost correct hypothesis of the nature of the image (the statistical stationarity and the locality of pixel dependency). Therefore, compared to standard feed-forward neural networks with similarly sized layers, CNN has far fewer connections and parameters and is therefore easier to train, and its theoretically best performance may only be slightly worse. Despite the attractive qualities of CNNs and the relative efficiency of their local architecture, it is still very expensive to apply them to high-resolution images on a large scale. Figure 18 summarizes the architecture of our network. It contains eight learning layers-five convolutional layers and three fully connected layers. As shown in figure 18, the network contains 8 weighted layers. The first five are convolutional, and the remaining three are fully connected. The output of the last fully connected layer is fed to a 1000-way softmax, which produces a distribution on 1000-class labels. Our network maximizes the objective of polynomial logistic regression, which is equivalent to maximizing the average value of the log probability training case of the correct label under the predicted distribution.

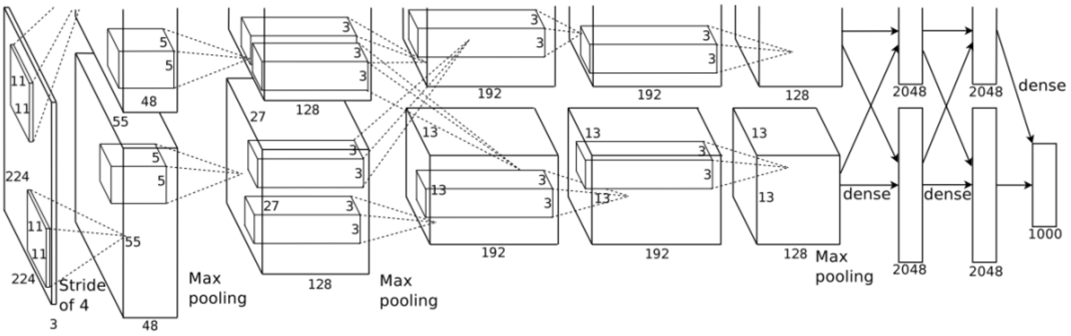


Figure 18 An illustration of the architecture of our CNN, explicitly showing the delineation of responsibilities between the two GPUs.

One GPU runs the layer-parts at the top of the figure while the other runs the layer-parts at the bottom. The GPUs communicate only at certain layers. The network's input is 150,528-dimensional, and the number of neurons in the network's remaining layers is given by 253,440–186,624–64,896–64,896–43,264– 4096–4096–1000.

The kernels of the second, fourth and fifth convolutional layers are only connected to those kernel maps in the previous layer on the same GPU (see Figure 18). The kernels of the third convolutional layer are connected to all kernel maps in the second layer. The neurons in the fully connected layer are connected to all neurons in the previous layer. The response normalization layer follows the first and second convolutional layers. The maximum pooling layer follows the response normalization layer and the fifth convolutional layer. ReLU nonlinearity is applied to the output of each convolution and fully connected layer. The first convolutional layer uses a 4-pixel step (this is the distance between the centers of the receptive fields of adjacent neurons in the kernel map) to filter the 224×224×3 input with 96 11×11×3 kernels image. The second convolutional layer takes the output of the first convolutional layer (response normalized and merged) as input and uses 256 kernels of size 5×5×48 to filter it. Third, fourth and fifth convolutional layers are connected to each other without any intermediate pooling or normalization layers. The third convolutional layer has 384 kernels of size 3×3×256, which are connected to the (normalized, merged) output of the second convolutional layer. The fourth convolutional layer has 384 kernels of 3×3×192 size, and the fifth convolutional layer has 256 kernels of 3×3×192 size. There are 4096 neurons in each fully connected layer.

As a large-scale deep convolutional neural network, ConvNets can use purely supervised learning to achieve record results on extremely challenging data sets. For tasks related to computer vision, including recognition, object detection, and semantic segmentation, engineering functions have been replaced by ConvNets computing functions. In addition to being able to express higher-level semantics, ConvNet is also more robust to scale changes, so it helps to recognize features calculated from a single input scale. But even with this robustness, pyramids are still needed to obtain the most accurate results. The main advantage of characterizing each level of the image pyramid is that it can generate multi-scale feature representations, where all levels (including high-resolution levels) are semantically powerful. In our research, due to changes in the proportion of urban objects, the image pyramid will be used as an effective feature extraction method.

4.3.2 Feature Pyramid Network (FPN)

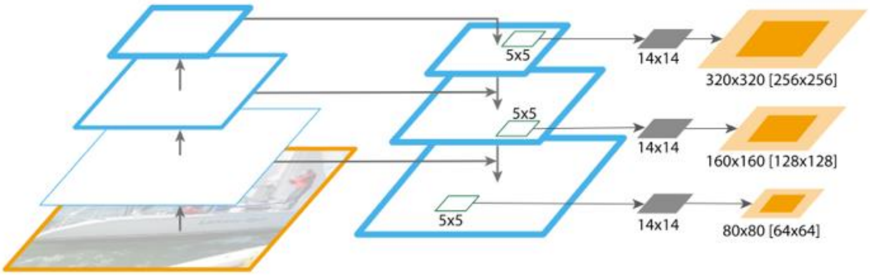


Figure 19 FPN Feature Extraction Architecture for Object Segment

For the building change detection problems, the greatest challenges consist of object differentiation including buildings, river, roads and moving objects, as well as the high density and scale difference of buildings. In the whole End-to-End model framework, the feature extraction network is required to extract object-based feature for urban objects with large scale divergence and segment building boundaries accurately and efficiently.

Feature pyramids built upon image pyramids form the basis of a standard solution. These pyramids are scale-invariant in the sense that an object's scale change is offset by shifting its level in the pyramid. Intuitively, this property enables a model to detect objects across a large range of scales by scanning the model over both positions and pyramid levels (Lin, 2017). FPN architecture and feature extraction process is shown in figure 19. Aiming at leveraging a ConvNet's pyramidal feature hierarchy, which has semantics from low to high levels, and build a feature pyramid with high-level semantics through-out. The Feature Pyramid Network is created general-purpose, so it could be applied flexibly for our building change detection problem without obvious limitations.

For network architectures, FPN usually takes single-scale images of any size as input, and outputs scale-scale feature maps at multiple levels in a fully convolutional manner. This process is independent of the backbone convolution architecture. The bottom-up path, the top-down path and the horizontal connection constitute the FPN pyramid. Figure 20 shows the building blocks of lateral connections and top-down paths. The use of FPN has the following advantages: the horizontal connection between the reconstruction layer and the corresponding feature map can help the detector to skip and predict the position connection more accurately, making training easier.

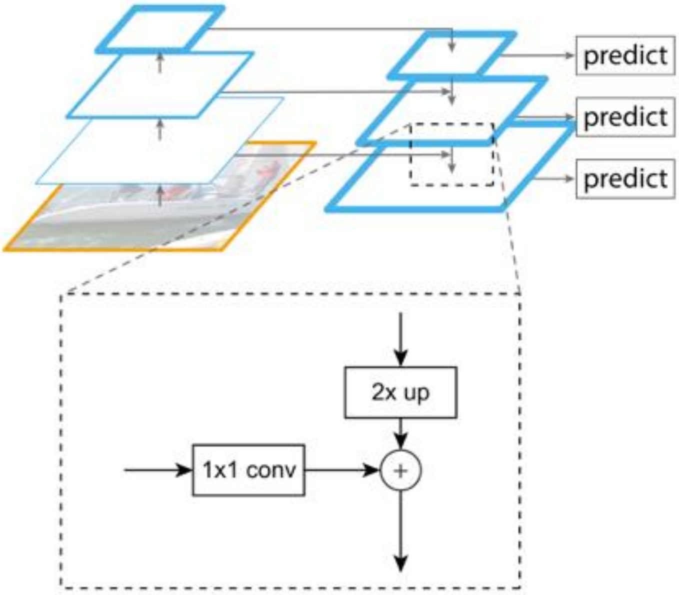


Figure 20 A building block illustrating the lateral connection and the top-down pathway, merged by addition.

4.4 Focal Loss for Multi-Class Classification

Classification prediction modeling involves predicting category labels for a given observation. An unbalanced classification problem is an example of a classification problem where the distribution of examples in a known class is skewed or skewed. The distribution can range from slight deviations to severe imbalances. There is one example in the minority, and hundreds, thousands, or millions of examples in the majority.

As addressed in Section 1 and Section 2, the sample number of continuations is approximately 3000 times higher than the sample number of new construction and demolishment, which causes a severe class imbalance in our multi-class classification problem. Imbalanced classification poses challenges to predictive modeling because most machine learning algorithms used for classification are designed around the assumption that each class assumes the same number of examples. This leads to poor prediction performance of the model, especially for minority groups. This is a problem, because in general, the minority category is more important, so the problem is more sensitive to the classification error of the minority category than the majority. For example, by default, the binary classification model is initialized to have an equal probability of output y = -1 or 1. Under this initialization, in the case of class imbalance, the loss due to frequently occurring classes may account for most of the total loss, and lead to instability in early training. In this way, except for sample adjustment on training dataset, Focal Loss is a principle approach we deployed for avoiding the unequal class distribution, and then promote the training efficiency of proposed end-to-end model.

Focal loss is a loss function, which can be used as a more effective alternative than traditional loss function methods for dealing with imbalance problems. The loss function is converted from the cross-entropy loss after dynamic scaling, where the scaling factor will decay to zero as the confidence in the correct category increases. The scaling factor can automatically reduce the contribution of simple examples during the training process, thereby effectively focusing the model on hard examples. Since Focal Loss may help to train a high-precision first-level detector, this method is significantly better than other training methods in terms of sampling heuristics or hard sample mining, so we compile the loss function in the network for training, thus Calculating the loss output of the change graph can also solve the problem of sample imbalance in our training data set.

The Focal Loss (Lin, 2018) was designed to address the one-stage object detection scenario in which there is an extreme imbalance between foreground and background classes during training (e.g., 1:1000). Lin introduced the focal loss starting from the cross entropy (CE) loss for binary classification:

$$CE(p,y) = \begin{cases} -\log(p) & if\ y = 1 \\ -\log(1-p) & otherwise \end{cases} \quad (5)$$

In the above $y \in \{\pm 1\}$ specifies the ground-truth class and $p \in [0, 1]$ is the model's estimated probability for the class with label $y = 1$. For notational convenience, we define p_t

$$p_t = \begin{cases} p & if\ y = 1 \\ 1-p & otherwise \end{cases} \quad (6)$$

A common method for addressing class imbalance is to introduce a weighting factor $\alpha \in [0, 1]$ for class 1 and $1-\alpha$ for class -1. In practice α may be set by inverse class frequency or treated as a hyperparameter to set by cross validation. The α-balanced CE loss is written as:

$$CE(p_t) = -\alpha_t \log(p_t) \quad (7)$$

However, the large types of imbalances encountered during the training of dense detectors overwhelm the loss of cross-entropy. Negative values that are easy to classify account for most of the loss and dominate the gradient. Although α balances the importance of positive/negative examples, it does not distinguish between simple/difficult examples. Instead, they suggest reshaping the loss function into a simple example of light weight, so that the training focus is on the hard negative. More formally, a modulating factor $(1-p_t)\gamma$ is added to the cross-entropy loss, with tunable focusing parameter $\gamma \geq 0$. And the focal loss is defined as:

$$FL(p_t) = -(1-p_t)^{\gamma}\log(p_t) \quad (8)$$

There two properties of the focal loss. (1) When an example is misclassified and p_t is small, the modulating factor is near 1 and the loss is unaffected. As $p_t \to 1$, the factor goes to 0 and the loss for well-classified examples is down-weighted. (2) The focusing parameter γ smoothly adjusts the rate at which easy examples are down- weighted. When $\gamma = 0$, FL is equivalent to CE, and as γ is increased the effect of the modulating factor is likewise in- creased (we found $\gamma = 2$ to work best in our experiments). Intuitively, the modulating factor reduces the loss contribution from easy examples and extends the range in which an example receives low loss. For instance, with $\gamma = 2$, an example classified with $p_t = 0.9$ would have 100× lower loss compared with CE and with $p_t \approx 0.968$ it would have 1000 × lower loss. This in turn increases the importance of correcting misclassified examples (whose loss is scaled down by at most 4 × for $p_t \leq 0.5$ and $\gamma = 2$). In practice, an α-balanced variant of the focal loss is used:

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \quad (9)$$

In our end-to-end change detection model, as a multi-class classification problem, a custom multi-class focal loss is used for change detection model with imbalanced datasets. α is defined as a 2-D array in the shape of (3,1).

$$a = [[\alpha_0], [\alpha_1], [\alpha_2]] \quad (10)$$

where $a_0, a_1, a_2 = weight\ value\ of\ each\ class$

To do element-wised loss calculation without α-balanced variant for will be written as:

$$FLE(\hat{y}, y^t) = \frac{\sum_{i=1}^{N} -(1-p_i^t)^{\gamma}\log(p_i^t)}{N} \quad (11)$$

where $\hat{y} =$ predicted result tensor,

$y^t =$ ground truth tensor

$$p^t = \begin{cases} p & if\ y = 1 \\ 1-p & otherwise \end{cases} \quad (12)$$

$N =$ tensor elements number

And the loss calculation we implemented is written as:

$$FLM(\hat{y}, y_{true}) = FLE(\hat{y}, y_{true}) * \alpha \quad (13)$$

In our experiments, the implementation of our custom loss layer combines the sigmoid operation for computing p with the loss computation, resulting in greater numerical stability. Meanwhile, the manual setting of multi α-balanced variant increases the flexibility and robustness of our loss function, in response to various sample imbalanced problem. In our model, it is only deployed to output3 layer for change map, which is a multi-classification network. At the start of training, the layer effectively counters the domination of the frequent class because of the 'privilege' for the rare function, finally making it even. As a result, the instantiation of proposed custom focal loss shows great improvement in our experiments, compared with traditional cross entropy, which will be specifically described in the following experiment Section.

Chapter 5

Experiment

Following the principle of ablation study, except for our proposed methodology, two baseline methods are conducted and executed simultaneously for the empirical results, performances of which are also illustrated in this Section. The experiment part will be given as follows: Section 5.1 describes the two baseline methods, algorithm and principles. Dataset introduction is given in Section 5.2. And Section 5.3 lists all used hyperparameters and the empirical results including visual ones as well as statistical measures and metrics.

5.1 Baselines

Due to the significant variation of datasets and tasks between our study and other scholars, it is not appropriate to execute baselines by realizing the algorithms of published scholars. Therefore, in order to evaluate the change indicators and validity of procedures, we implement image differencing post-classification method and similar dual end-to-end framework with only 2-D imagery, which is aiming at validating the significance of end-to-end model method and the introduction of the 3-D information in DSM. Following the methodology of Ablation Study, the empirical study will result in convincing the function of additional components, methods and datasets.

5.1.1 Post-Classification Method

Baseline1 could be summarized as a hybrid pixel-based method combining post-classification with segmentation model (see figure 21). The implementation could be concluded as following steps:

1. Multi-temporal ortho aerial image of T1(Time 1) and T2(Time 2), without DSM but only spectral information, will be inputted into FPN building semantic segmentation network after pre-processing consisting of SIFT image registration and image radiometric correction. The FPN building semantic segmentation network will be trained by the dataset including both 2015 and 2016 dataset. Building segmentation maps of 2015 and 2016 will be outputs of this step.

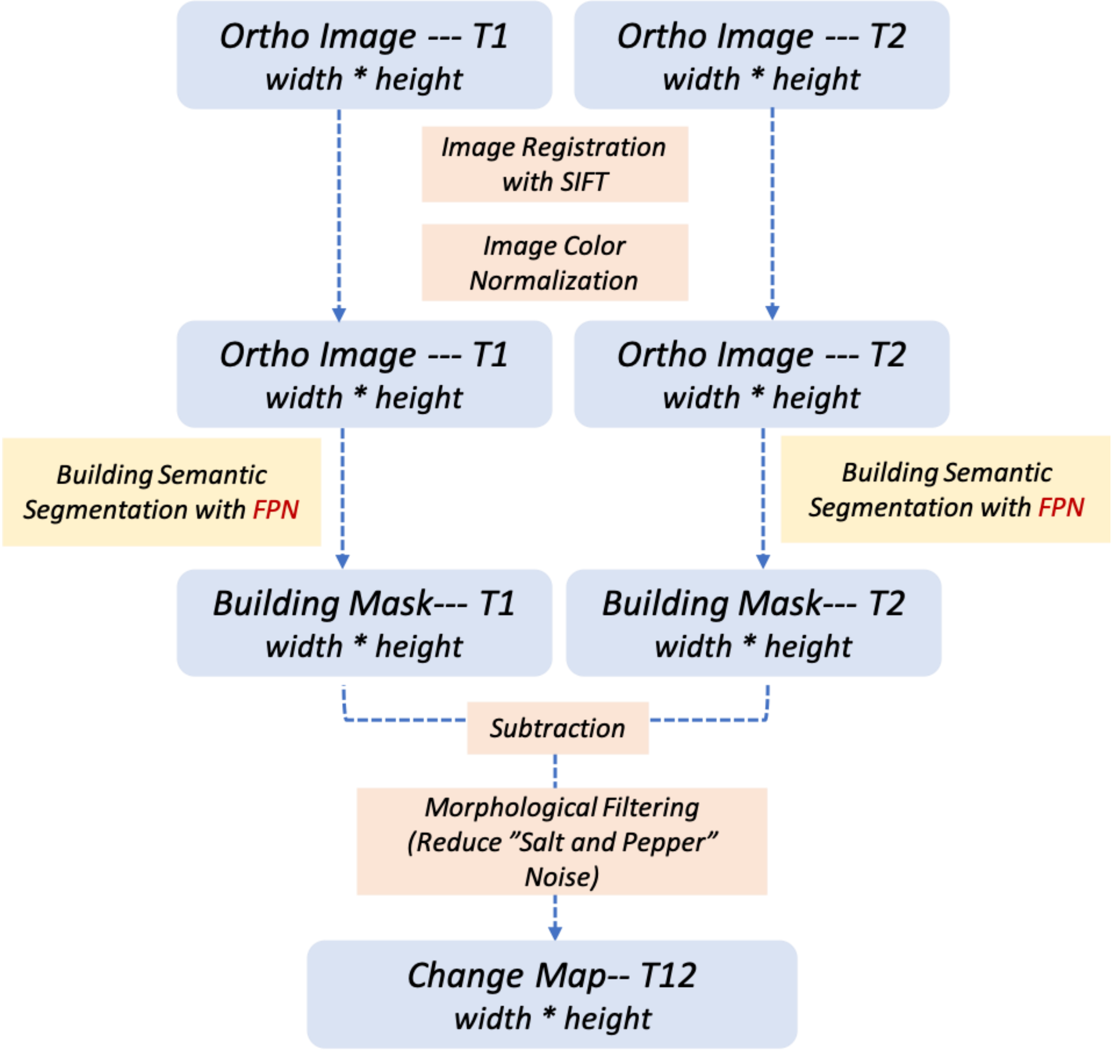


Figure 21 Framework of Post-Classification Methods as Baseline 1

2. Then, the two outputted building masks will take element-wised subtraction operation. Specifically, the binary-classification task in last step will generate segmentation map, in which every pixel will be classified into 'foreground' and 'background', represented by 1 and 0. As a mathematical subtraction result, the -1, 0 and 1 at each corresponding position will be represented by (0, 0, 0), (128, 128, 128) and (255, 255, 255) in output subtraction map.

3. However, pixel-based methods all have a problem called salt and pepper effect which means independent pixels that are classified wrong will cause bad performance of the whole map, even the accuracy is not low. It is a form of noise sometimes seen on images, also known as impulse noise. This noise can be caused by sharp and sudden disturbances in the image signal. It presents itself as sparsely occurring white and black pixels. Therefore, the morphological filtering algorithm will be employed to reduce noise for the intermediate result then we will get the final change map.

Morphological Filtering basically consists of non-linear operations, which may be related to the shape or shape of image features, such as boundaries, skeletons, etc. In morphological technology, the detection image by masking a small template or shape called a structural element defines an area of interest or neighborhood around the pixel based on this element. Mathematical morphology is based on set theory operations, which are defined between a set of points in an image called an object and a core called a structural element. These are some basic morphological operations: a) Dilation b) Erosion c) Opening d) Closing e) Hit or miss Transform. In our baseline1, according to our denoising requirements, opening operation is utilized flexibly, in which dilation and erosion are both applicated

Morphological dilation is an operation that involves finding the maximum value among the pixels belonging to the window. Dilation eliminates noisy pixels that appear in the object area as the object size increases. Basically, it adds pixels at the boundary of the object in the image, which means that if the image is expanded, the area of the object area will increase. There are two operands: input image and structured element (SE). Enter here an image I of size $G \times H$ and a structural element B of size $K \times L$, which define the size of the window. Mathematically, it can be written as:

$$[I \oplus B](w,l) = \max[I(w-u, l-u) \mid (u,v) \in B] \quad (14)$$

In other terms, dilation can be expressed as:

$$[I \oplus B] = [m \in Z^2 \mid m = i + b, i \in I, b \in B] \quad (15)$$

where m is a set of points. So, dilation is an increase in pixels to the boundary of the object in the image.

Morphological erosion refers to the operation of finding the minimum value among the pixels belonging to the window. Erosion removes noisy pixels present in the background and reduces the size of the object. Basically, if the image is eroded, the pixels at the object boundary are subtracted from the image mean, and the area of the object area is reduced. There are two operands: input image and structured element (SE). Enter image I of size $G \times H$ and B of size $K \times L$ (define the size of the window) here. Mathematically, it can be written as:

$$[I \ominus B](w,l) = \min[I(w-u, l-u) \mid (u,v) \in B] \quad (16)$$

In other terms, Erosion can be expressed as:

$$[I \ominus B] = [m \in Z^2 \mid m = i + b, i \in I, b \in B] \quad (17)$$

where m is a set of point. So, erosion is nothing but subtraction of pixels to the boundaries of object in the image.

The morphological opening on the image is defined as B eroding I, and then using B to expand the eroded image. Mathematically, it can be expressed as:

$$[I\ o\ B] = (I\ \ominus B)\ \oplus B \quad (18)$$

Through the opening operation, the external noise existing in the background area is eliminated, and the object remains because it is original. In this way, the final change map will be generated and show better performance after salt-and-pepper effect being eliminated, than the subtraction map. The general effect of morphological opening is shown in figure 22.

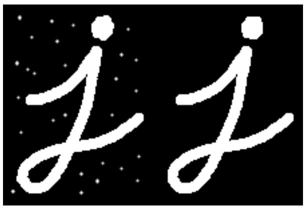


Figure 22 General Effect of Morphological Opening

The framework of baseline1 is shown in Figure 21. The biggest advantage of this method lies in the employing of existing advanced segmentation network to accurately extract building features, which take use of current advanced techniques and promote the output efficiency. Nevertheless, manual fine-tuning of thresholds and pixel-wised subtraction could not guarantee the robustness of this method. The next baseline uses the same data and feature extraction network FPN, but different architecture, as an end-to-end model is established and applicated.

5.1.2 Dual-Input End-to-End Model

Baseline 2 is an end-to-end model-based method with multi inputs and multi outputs, combining FPN and Focal Loss. In this model, simultaneously, a binary-classification building semantic segmentation problem and a multi-classification building change detection problem are trained together in an end-to-end model.

As for the whole network architecture, multi-temporal ortho aerial image of T1 and T2, without DSM, will be inputted into the FPN-based model after same pre-processing process. The model architecture is similar with our proposed method (described in Section3), except for the DSM input and the concatenate layer of two difference maps. Specifically, the whole model comprises FPN architecture, one subtraction layer and one convolutional block consisting of three convolutional layers with (3,3) kernel size and 32 channels. All the layers share parameters for from two input pipelines and be trained simultaneously. For the optimizer of this model, the overall loss includes two binary cross entropy losses between intermediate outputs and building segmentation ground truth as well as the focal loss between final output and change map ground truth. Loss weights are distributed equally as [1,1,1]. The general framework of baseline 2 is shown in Figure 23. The

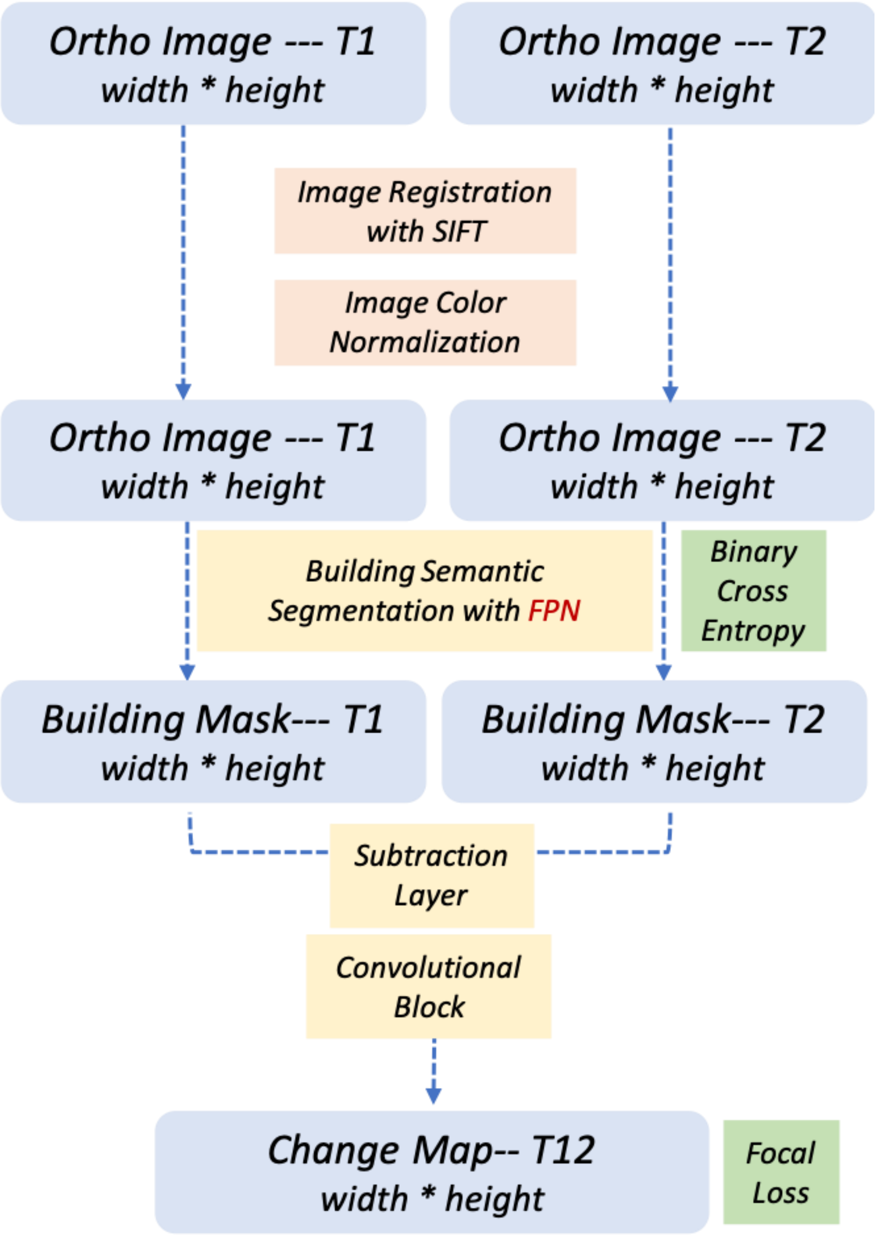


Figure 23 Framework of Dual-Input End-to-End Model Method as Baseline2

framework of shape transformation processing and feature composition is shown in figure 24. As the segmentation is binary classification problem, the generator will make them tensors in the shape of $(N_{patch}, 224, 224, 2)$ with one-hot encoding and the change map ground truth will be transformed to $(N_{patch}, 224, 224, 3)$ with 3 classes. The intermediate output from FPN will be two binary classified tensors with shape $(N_{patch}, 224, 224, 2)$ and value in $\{0,1\}$. Later a subtraction layer merges these two tensors with element-wised subtraction and get one tensor with $(N_{patch}, 224, 224, 2)$ shape and value in $\{-1, 0,1\}$. It will go through a convolutional block comprises 3 convolutional layers with (3,3) kernel, 32 channels and same padding, then output change map tensor in $(N_{patch}, 224, 224, 3)$ as final output. The spatial size as well as resolution will remain same with original inputs.

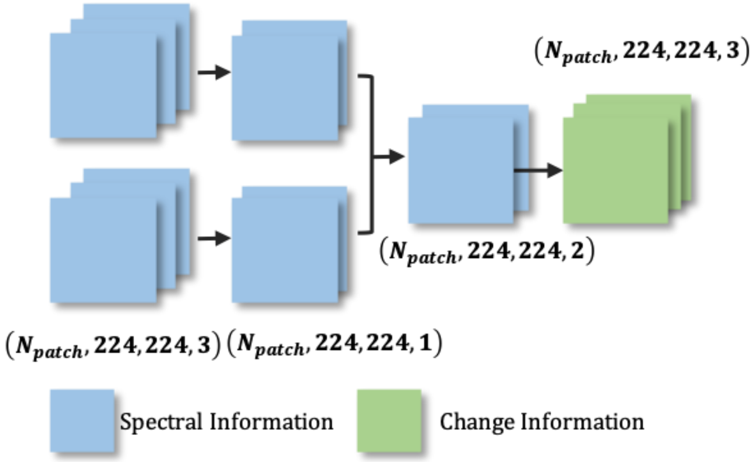

$(N_{patch}, 224, 224, 3)$
$(N_{patch}, 224, 224, 2)$
$(N_{patch}, 224, 224, 3)$
$(N_{patch}, 224, 224, 1)$
Spectral Information
Change Information

Figure 24 Shape transformation and feature composition of Baseline2

5.2 Data Description

The dataset used for all experiments conducted is Tokyo aerial imagery dataset provides by NTT Spatial Information Company, Japan. For experimental results verifying the effectiveness and generality of our proposed framework, we used the dataset covering the area of Setagaya-Ku, Tokyo, Japan, consisting of 15 ortho imagery grids that acquires 2000m×1500m field size, 12500×9375 pixels and 0.16m resolution as well as 1050 raw aerial imagery with sufficient overlap rate over 70%.

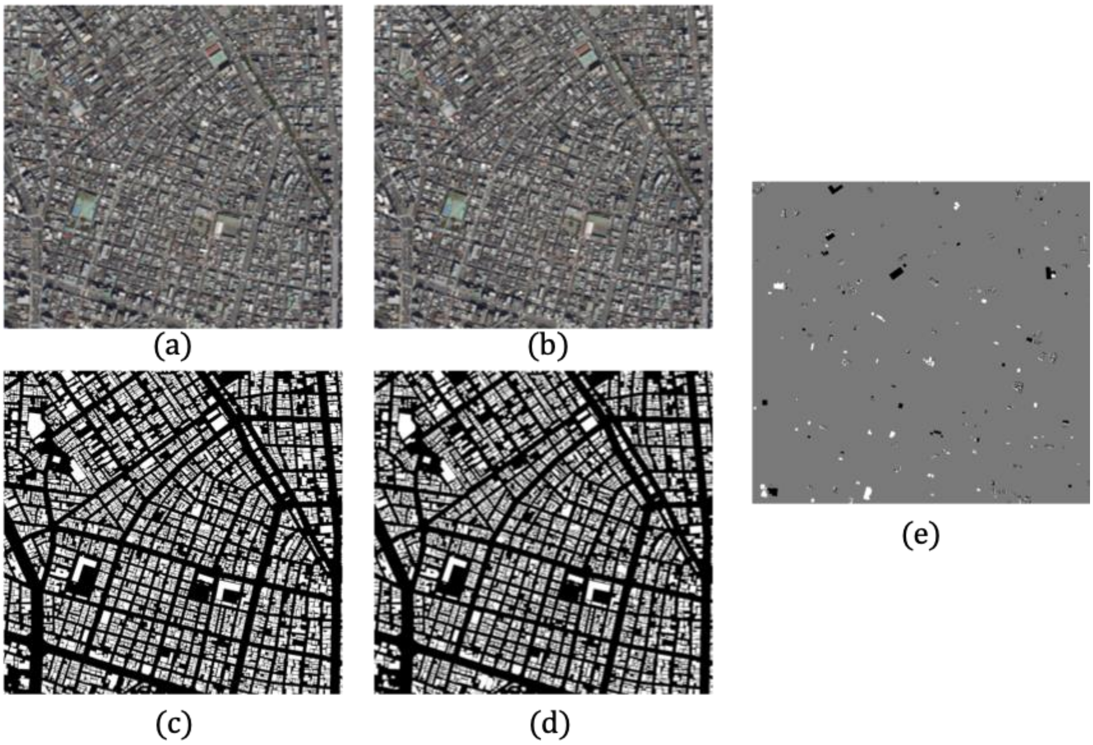


Figure 25 Training Dataset Composition Examples

(a) Ortho Image taken in 2015 (b) Ortho Image taken in 2016 (c) Building Segmentation Ground Truth of 2015 (d) Building Segmentation Ground Truth of 2016 (e) Change Map Ground Truth

The building segmentation ground truth is binary-labelled per-pixel to foreground and background classes representing building label and no-building label. And the change map ground truth is classified into three labels representing building new construction, construction continuation and building demolishment. In summary, the training dataset is made up for 2015 and 2016 datasets, each of which consists of 2 ortho aerial images, corresponding building segmentation labels ground truth, 5 DSM images in the same size, as well as change map ground truth with same width and height, as shown in figure 25. The area we use for testing is part of Arakawa-Ku which is a typical urban area with constructions, road networks, rivers and other urban infrastructures.

All the training and testing data are clipped into small patches in 224×224 pixels in order to reduce the random data noise, improve the model compatibility and make flexibility for data augmentation. All the DSM data is normalized to the range [-1,1] and the ground truth labels will be transferred to one-hot encoding. In the training process, the validation set occupies 20% of whole training set and stay still in every epoch for comparison.

For the output tensor, the value in each channel of each output pixel thus represented the confidence of the model in classifying the pixel as belonging to a specific class. In order to determine the class that the model finds most likely for a pixel, the argmax function, which returns the class with the highest confidence value, was applied to the output tensor.

5.3 Empirical Results

In order to evaluate the performance of all the methods quantificationally, we use evaluation metrics to compare the change difference images with ground-truth maps, in which white pixels represent new construction, black pixels mean demolishment and grey pixels means continuation. Generally, through pixel-level evaluation, this paper adopts the following five evaluation criteria: confusion matrix, overall accuracy (OA), precision, recall, F1 score. In their calculation, there are four indexes: 1) TP: true positives, i.e., the number of correctly detected changed pixels; 2) TN: true negatives, i.e., the number of correctly detected unchanged pixels; 3) FP: false positives, i.e., the number of false-alarm pixels; and 4) FN: the false negatives, i.e., the number of missed changed pixels.

The Confusion Matrix comprehensively describes the performance of the classification result for every class:

True Negatives (TN)	False Positives (FP)
False Negatives (FN)	True Positives (TP)

The OA is defined as:

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \quad (19)$$

The Precision is defined as:

$$Precision = \frac{TP}{TP + FP} \quad (20)$$

The Recall is defined as:

$$Recall = \frac{TP}{TP + FN} \quad (21)$$

The F1 score is defined as:

$$F1 = 2 * \frac{Recall * Precision}{Recall + Precision} \quad (22)$$

The performance of baseline1, baseline2 and our proposed model are demonstrated in Section 5.3.1, Section 5.3.2 and Section 5.3.3 respectively. The results of these three algorithms are compared and discussed in Section 5.3.4.

Given that the baselines and proposed method are all model-based, the model initial settings and hyperparameters will be described as followed. For the implementation of three models of above methods, we adapt Adam optimizer with learning rate as 0.00011 and momentum ratio as 0.9, with batch size as 32 and training epoch number as 200. Early stopping mechanism was set as 10 epoch patience from validation loss value. Checkpoint was also set according to validation loss to save optimal model weights. Relu is used as activation function in convolutional layers and Softmax function is used for output layers.

5.3.1 Results of Baseline 1

In Baseline1, the model is only for semantic segmentation task. As a post-classification method, the model was trained and tested with 2015 dataset and 2016 dataset separately. Binary cross entropy was the only loss function employed to segmentation output. In the morphological filtering part, opening operation was implemented to reduce salt-and-pepper noises with erosion kernel size and dilation kernel size set as 15 and 20 respectively. Figure 26 illustrates the testing result of post-classification method in baseline1, with two typical area in Tokyo as testing data.

For single-input and single-output FPN for segmentation, at encoding/down-sampling stage, all convolution operations in the model had a kernel size of 3×3, a stride of 1, and batch normalization. On the top of bottom-up pathway, kernel being applied to reduce channels is in size 1x1.Up-sampling operations had kernel sizes including (8,8), (4,4) and (2,2), as well as the nearest-neighbor interpolation method. And the convolutional layer for transferring to feature map is in size 3×3. The input and output tensor dimensionality are (2355, 224, 224, C) where C means channel number differentiates according to class number and data type.

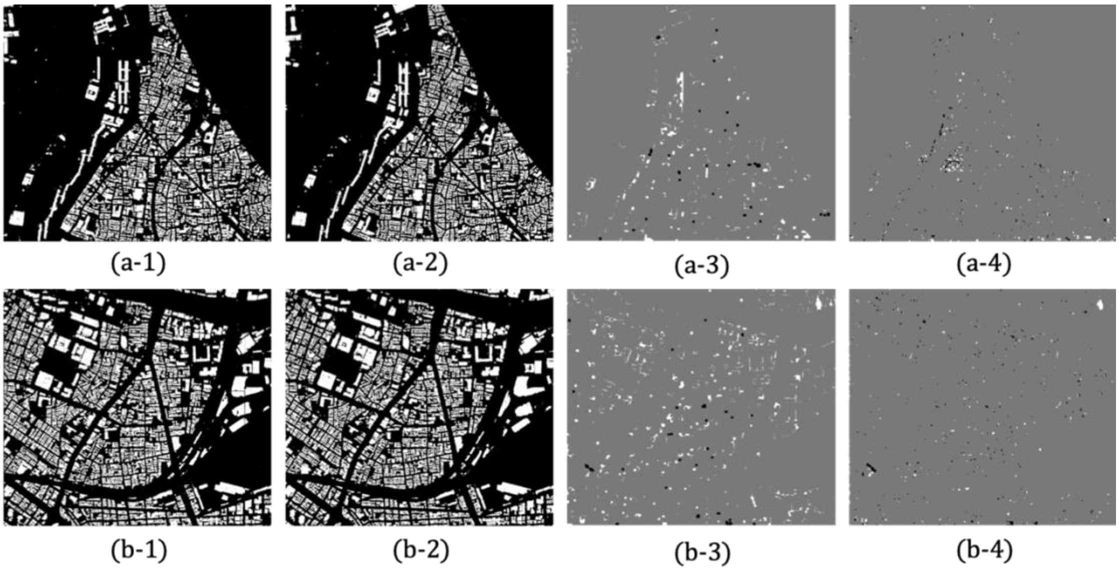


Figure 26 Testing results of Baseline1(Post-Classification Methods)

(a-1) Building Segmentation Map of area-A in 2015 (a-2) Building Segmentation Map of area-A in 2016 (a-3) 2015-2016 Change Map of area-A without Morphological Filtering (a-4) 2015-2016 Change Map of area-A with Morphological Filtering (b-1) Building Segmentation Map of area-B in 2015 (b-2) Building Segmentation Map of area-B in 2016 (b-3) 2015-2016 Change Map of area-B without Morphological Filtering (b-4) 2015-2016 Change Map of area-B with Morphological Filtering

From figure 27, we could find out that FPN shows excellent performance in building semantic segmentation task, with clear boundary and low noise. Nevertheless, the direct subtraction of multi-temporal building segmentation map will cause salt-and-pepper effect, including independent pixels and building outlines as noises. Although the image registration has been implemented in pre-processing procedure, mis-alignment still exists and effected the performances. With morphological opening operation, it shows that the denoising operation makes significant improvement in change detection performances. In the training process of the building semantic segmentation part, multiple networks are implemented and experimented, including FCN8s, FCN32s. U-Net and FPN, the corresponding OA of final change map is shown in table 1., with other parameters illustrated.

Table 1 OA of Baseline 1

Segmentation Network	Iteration (Early Stopping)	Morphological Opening Kernel Size	**OA (Overall Accuracy)**
FCN-8S	32	15,20	85.17%
FCN-32S	56	15,20	82.54%
U-Net	52	15,20	89.65%
FPN	47	15,20	91.96%

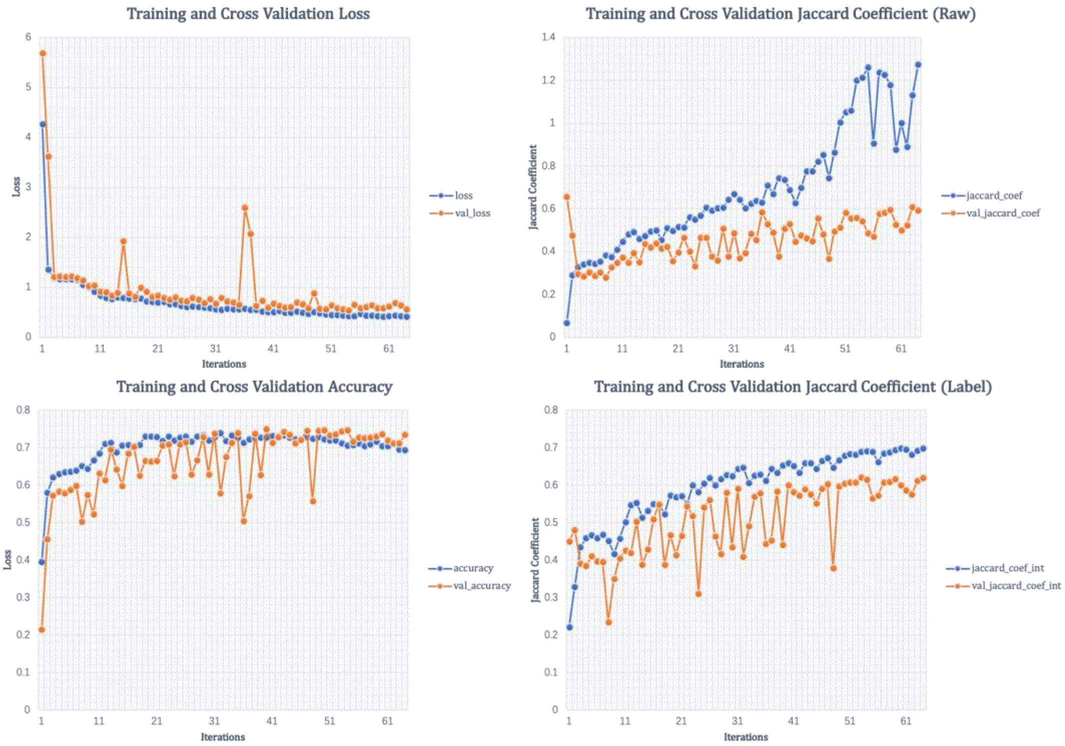


Figure 27 Learning Curves of Segmentation Network

Figure 27 comprises: (1) Training and Cross Validation Loss (2) Training and Cross Validation Accuracy (3) Training and Cross Validation Jaccard Coefficient (Raw) (4) Training and Cross Validation Jaccard Coefficient (Label). From figure 27, the training and corresponding validation process is demonstrated, we would see the tendency of model learning ascends rapidly in the beginning and stably afterwards. The early stopping mechanism was introduced, and the training finished at epoch 63. Although there are occasionally unnormal result and noise occurring, the uptrend of training and validation basically keeps consistent, which means the supervised learning was efficient and stable. The efficiency of FPN network architecture for semantic segmentation task was also certified.

5.3.2 Results of Baseline 2

In baseline2, as an end-to-end model method, it was trained with 2015 and 2016 dataset together, sharing parameters in convolutional blocks. All concatenate operations were implemented by axis 3 and subtraction layers are layer merging operation with no parameters trained. For multi-input and multi-output FPN, at encoding/down-sampling stage, all convolution operations in the model had a kernel size of 3×3 and batch normalization. On the top of bottom-up pathway, kernel being applied to reduce channels is in size 1x1. And the convolutional layer for transferring to feature map is in size 3×3. The input and output tensor dimensionality are (4710, 224, 224, C) where C means channel number differentiates by class number and data type. As the training process of 2015 and 2016 data is executed separately, the visualization result will be shown in Section 5.3.4, in comparison with other methods.

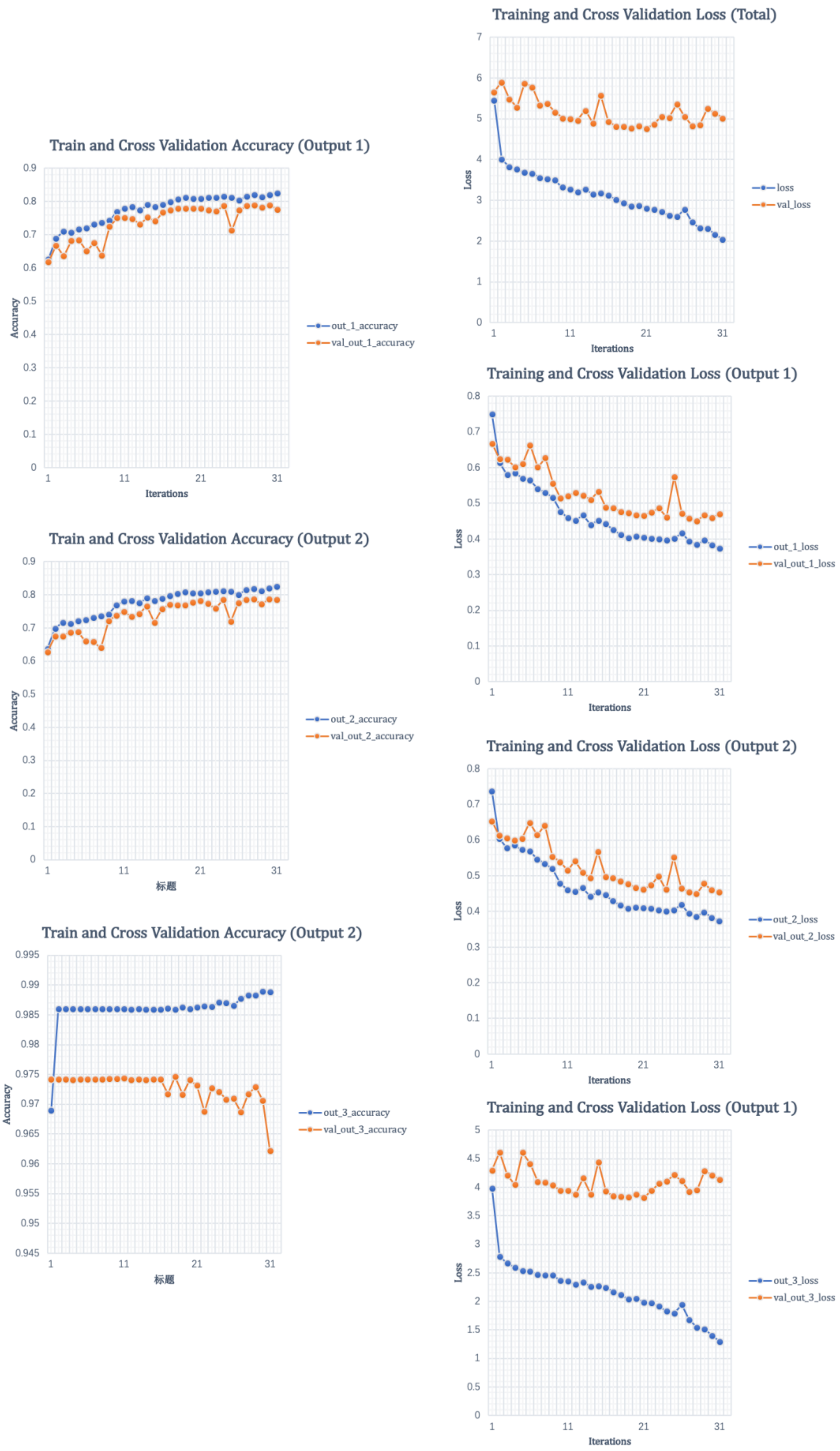


Figure 28 Learning Curves of Baseline2 for Losses and Accuracy

The learning curve of baseline two is illustrated in figure 28, the total learning process is efficient, when it is displayed in loss curve and accuracy curve. When it is separated to output1, output 2 and output 3, the learning processes of two segmentation outputs are similar. However, it has not been fully trained because the early-stopping mechanism stopped the training at epoch 34, because the validation loss has not descended in 10 epochs. The total loss was mainly dominated by loss of output3, which has rare improvement space. The output3 loss descends rapidly in the beginning and remain small fluctuation in the following. Another problem is that the validation loss is always much higher than training loss. It is crucial to conduct solutions such as loss distribution, facing this problem.

5.3.3 Results of Proposed Method

In our proposed method, as discussed in Section3, four inputs are received, and three outputs are generated. Our proposed end-to-end model was also trained with 2015 and 2016 dataset including imagery and DSM together. All the imagery are also clipped into patches in size 224 × 224, including DSM with one channel. All concatenate operations were implemented by axis 3 and all subtraction layers are layer merging operation with no parameters trained.

Same as above, the learning curve of our proposed method is shown in figure 29. The early stopping mechanism is also carried out and it finally stopped at epoch 38. With 4 inputs and 3 outputs, the introduction of DSM provided auxiliary information for the network training. We could find out that the problem of segmentation inadequate training and fluctuation of output3 loss still exist. However, it is delightful that the validation loss is in the similar tendency with training loss, with value difference much smaller than baseline2. The training efficiency has been promoted, which will consistently lead to better performance of the model.

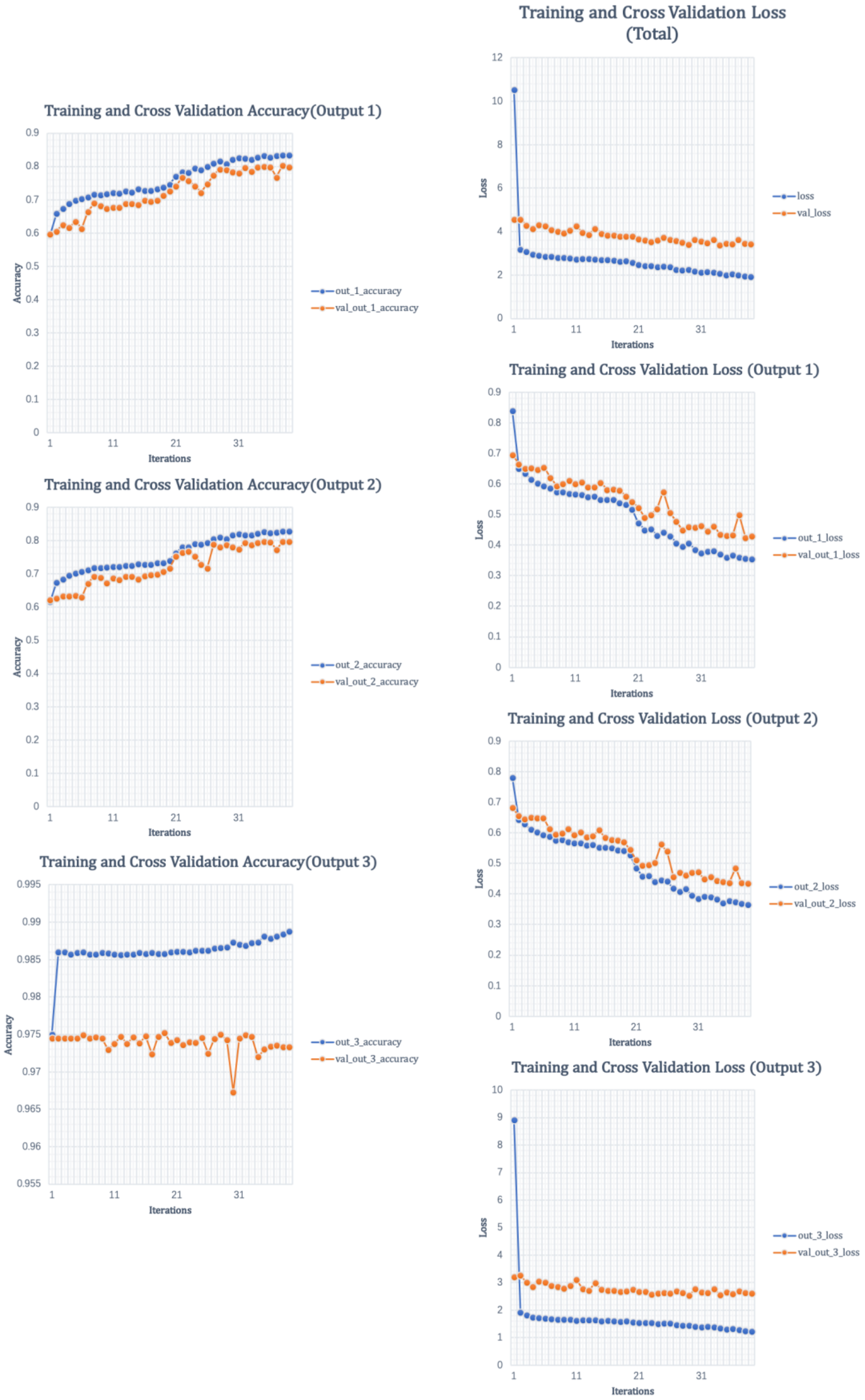


Figure 29 Learning Curves for Proposed Method of Losses and Accuracy

5.3.4 Comparison Results

Figure 30 shows part of experiments testing results of baselines and our proposed method, including building semantic segmentation results, change detection results and corresponding comparison to ground truth. In the fourth column, change map comparison color map of three methods are displayed in which blue represents true prediction and red represents false.

As shown in figure 30, for the building semantic segmentation results, the baseline1 had much better performance than the other two because of the model loss concentrating on the segmentation task only. The optimizer tends to optimize this binary classification for higher accuracy and lower noise. Nevertheless, the performance of post classification methods heavily relies on the capability of segmentation model and much noises arise due to the misalignment even though the segmentation has high accuracy, which need to be eliminated by morphological filtering.

Our method behaves relatively better than the baseline2 in segmentation results. Despite the two end-to-end models also distribute the optimizing concentration to change map by hyper losses, the DSM introduction also contribute to exact construction features by importing elevation details information.

For the change detection results, by visual inspection, the result of baselin1 generates severe noises due to the sample misalignment and inadequate morphological filtering in spite of the pre-processing and manual fine-tuning. Two end-to-end models have much better performances and our proposed method increased the ratio of True Positive and True Negative.

From quantitative analysis results shown in Table1., after overall consideration of OA, precision, recall and F1 score, it could be obviously found out that our proposed 4-input end-to-end model with imagery and DSM is an effective method for building change detection. Besides, the overall situation is that imbalanced dataset problem is severe that overall accuracy shows quite good performance, but other measurements calculated for all classes separately evaluate models worse. The post classification method got severe false positive samples situation and model with only imagery inputs got relatively low TP, which means it was not strong enough to have optimal prediction performance.

Table 2 Evaluation Metrics of Change Detection Baselines and Proposed Method

	Post-Classification	End-to-End Model(2-D)	End-to-End Model(3-D)
OA	0.954	0.960	0.971
Precision	0.479	0.634	0.680
Recall	0.371	0.520	0.511
F1 Score	0.418	0.571	0.583

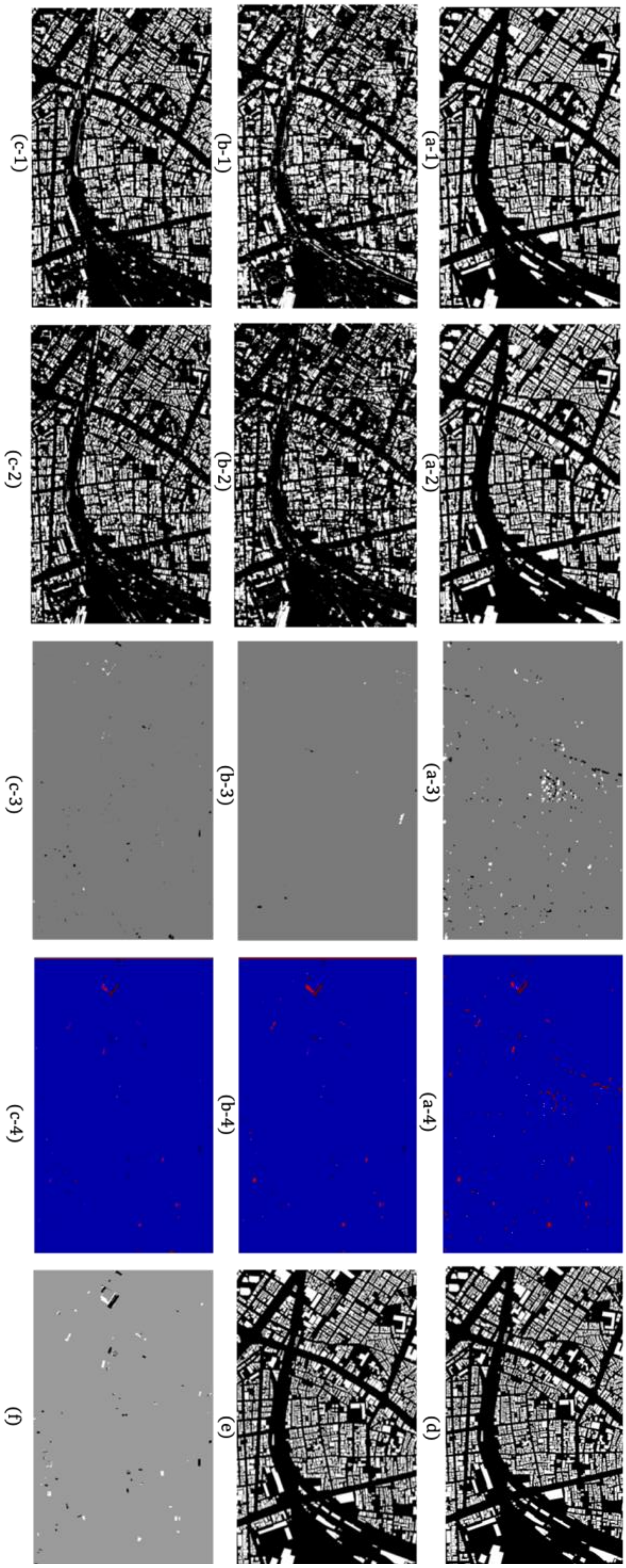


Figure 30 Testing Results of Comparison Experiments

Figure 30 consists of: (a-1) 2015 Segmentation Result of Baselin1 (a-2) 2016 Segmentation Result of Baselin1 (a-3) Change Map Result of Baseline1(a-4) Change Map Visual Evaluation of Baseline2 (b-1) 2015 Segmentation Result of Baselin2 (b-2) 2016 Segmentation Result of Baselin2 (b-3) Change Detection Result of Baseline2 (b-4) Change Map Visual Evaluation of Baseline2 (c-1) 2015 Segmentation Result of Proposed Method (c-2) 2016 Segmentation Result of Proposed Method (c-3) Change Detection Result of Proposed Method (c-4) Change Map Visual Evaluation of Proposed Method (d) 2015 Segmentation Map Ground Truth (d) 2016 Segmentation Map Ground Truth (d) Change Map Ground Truth

Chapter 6

Conclusions and Future Work

6.1 Conclusions and Discussions

In this dissertation, we propose a supervised end-to-end deep learning method based on Convolutional Neural Networks and Feature Pyramid Networks for multi-temporal building change detection in urban area via high-resolution aerial imagery and Digital Surface Models. The corresponding deep learning methods help us to train an end-to-end model that could execute classification tasks for both building semantic segmentation and change detection, with significantly prominent performance and high accuracy. Utilizing the acquired the end-to-end model-based classifier, with urban spectral, height and timely sequence information extracted from remote sensing, the segmentation maps and change maps could be generated automatically.

As a novel empirical study, the result of our proposed method is experimentally demonstrated together with baselines, during which the training and validation is executed on the aerial imagery dataset of Setagaya-Ku, Tokyo. As a high-density typical metropolitan area, there are diversified urban texture in Tokyo. Especially in Setagaya-Ku, there are rich natural landscapes, rivers, parks and chronological communities, offering divergent samples for model optimization. Also, the efficiency of applicated FPN network architecture, end-to-end argument, DSM introduction and other components are all validated with baselines, following the theory of ablation study.

Our proposed model mainly offers the following contributions. First, this supervised end-to-end method eliminates costly manual finetuning of thresholds and parameters. The skipping connection structure of FPN also decreases parameter amount and increases training effectiveness. Second, adaptive image pre-processing step and weight distribution of model optimize the change performance and minimize the noise simultaneously. Third, DSM introduction helps to extract deeper imagery feature for identification and classification.

Furthermore, our proposed method can be efficiently utilized as diffusely practical application in remote sensing multi-temporal building change detection in urban environment. By training the prior knowledge of the corresponding samples of building semantic segmentation and change detection, the proposed method could generate a classifier that acquires the capability to classify relative targets and leads to authentic classification results. Therefore, based on the empirical results of the experiments, as well as the low operational difficulty of end-to-end method, our proposed end-to-end FPN-based change detection model is a promising approach that has the potential to be widely applied to urban scenarios including government administration, real estate market, urban planning and so on.

6.2 Future Work

Although this study indicates the proficiency, efficiency and accuracy of our proposed model for building change detection, further and more detailed exploration on the development and expansion of this methods is still required in the future.

Initially, the method is based on both remote sensing aerial imagery and DSM in 0.16m resolution, which makes it difficult to prepare dataset and realize real-time detection for potential practical system. For the resolution of this basic problem, more related algorithms should be further developed, including color normalization and super resolution, by which a more complicated, comprehensive and robust system could be utilized in greater application scenarios. Second, as a model-based method on supervised learning, it is necessary to alleviate the labor-intensive training data labeling task when being applied to other remote sensing even computer vision dataset. Nonetheless, the difference in image capture condition and camera category will lead to the huge difference in illumination, perspective and sharpness of imagery, not to mention the following effect on Photogrammetry to generate ortho imagery, DTM and DSM. Therefore, to improve the performance in such cases, appropriate transfer learning technique should be carried out, aiming at higher universality and robustness of our method. Third, our proposed model is a multi-task classifier that could execute building semantic segmentation and change detection simultaneously. However, from Section 5, it demonstrated that the performance of segmentation does not surpass that of baseline1 or baseline2. Considering the loss combination way and weight distribution, current architecture could only concentrate on one classifier and guarantee the normal but not optimal performance of others. Further, the modification of networks architecture should be taken into count, because of the fact that optimal feature extraction network could produce more accurate and comprehensive semantics, for both building identification task and change detection task. Last but not least, we will apply the proposed method to other sequence feature identifications and raise its sensitivity for spatial-temporal data. In the future, not only the three class: continuation, demolishment and new-construction, but other labels including: re-construction, function change or residents change could be detected as well. We believe in its great potential in practical value.

Beyond that, there are still many applications could be researched and developed with our existing dataset, such as Object-based Urban Environment Change Prediction with Multiple Multi-temporal Remote Sensing Imagery. As shown in figure 31, remote sensing imagery and DSM of the same area taken over 10 years are used as training data. The model will compare and learn the timely construction behavior pattern in this area, and generate a prediction change map of corresponding area construction change situation in the coming year automatically. This prediction result could also be significant working for government management and real estate market as well.

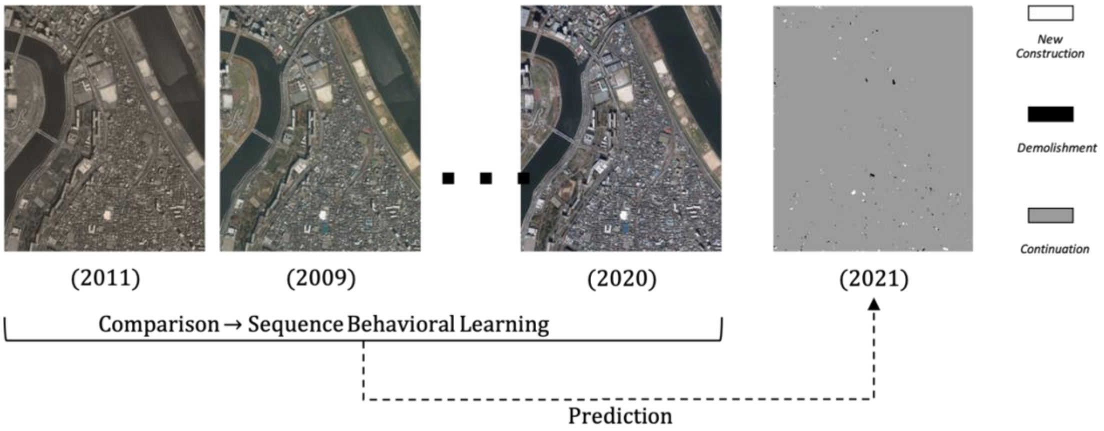


Figure 31 Future Research: Urban Environment Change Prediction

In addition to all above, newly published papers on computer vision, machine learning, deep learning and remote sensing in the future could also inspire us on how to further optimize our method. Relevant domains such as urban computing and smart transportation are also possible to be combined with our method, implementing further researches which is also aiming at the promotion of urban society, becoming more smart, convenient and prosperous

Bibliography

[1] Singh, A. (1989). Review Articlel: Digital change detection techniques using remotely-sensed data. International Journal of Remote Sensing, 10(6), 989–1003.

[2] Hussain, M., Chen, D., Cheng, A., Wei, H., & Stanley, D. (2013). Change detection from remotely sensed images: From pixel-based to object-based approaches. ISPRS Journal of Photogrammetry and Remote Sensing, 80, 91–106.

[3] Lu, D., Mausel, P., Brondízio, E., Moran, E. (2004) Change detection techniques: International Journal of Remote Sensing, 25 (12), pp. 2365-2401.

[4] Roysam, B., Al-Kofahi, O., Radke, R. J., & Andra, S. (2005). Image change detection algorithms: A systematic survey. IEEE Transactions on Image Processing, 14(3), 294–307.

[5] Gong, J., Sui, H., Ma, G., & Zhou, Q. (2008). A review of multi-temporal remote sensing data change detection algorithms. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives, 37, 757–762.

[6] Wang, G. H., Wang, H. Bin, Fan, W. F., Liu, Y., & Liu, H. J. (2018). Change detection in high-resolution remote sensing images using levene-test and fuzzy evaluation. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives, 42(3), 1695–1701.

[7] Mou, L., Bruzzone, L., & Zhu, X. X. (2019). Learning spectral-spatialoral features via a recurrent convolutional neural network for change detection in multispectral imagery. IEEE Transactions on Geoscience and Remote Sensing, 57(2), 924–935.

[8] Tian, J., Cui, S., & Reinartz, P. (2014). Building change detection based on satellite stereo imagery and digital surface models. IEEE Transactions on Geoscience and Remote Sensing, 52(1), 406–417.

[9] Qin, R., Tian, J., & Reinartz, P. (2016). 3D change detection – Approaches and applications. ISPRS Journal of Photogrammetry and Remote Sensing, 122, 41–56.

[10] Gong, M., Zhan, T., Zhang, P., & Miao, Q. (2017). Superpixel-based difference representation learning for change detection in multispectral remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 55(5), 2658–2673.

[11] Xiao, P., Zhang, X., Wang, D., Yuan, M., Feng, X., & Kelly, M. (2016). Change detection of built-up land: A framework of combining pixel-based detection and object-based recognition. ISPRS Journal of Photogrammetry and Remote Sensing, 119, 402–414.

[12] Zhang, X.; Shi, W.; Lv, Z.; Peng, F. Land Cover Change Detection from High-Resolution Remote Sensing Imagery Using Multitemporal Deep Feature Collaborative Learning and a Semi-supervised Chan–Vese Model. Remote Sens. 2019, 11, 2787.

[13] De Jong, K. L., & Sergeevna Bosman, A. (2019). Unsupervised Change Detection in Satellite Images Using Convolutional Neural Networks. Proceedings of the International Joint Conference on Neural Networks, 2019

[14] Wang, Q., Yuan, Z., Du, Q., & Li, X. (2019). GETNET: A General End-To-End 2-D CNN Framework for Hyperspectral Image Change Detection. IEEE Transactions on Geoscience and Remote Sensing, 57(1), 3–13.

[15] Peng, D., Zhang, Y., & Guan, H. (2019). End-to-end change detection for high resolution satellite images using improved UNet++. Remote Sensing, 11(11).

[16] B. C. Matei, H. S. Sawhney, S. Samarasekera, J. Kim and R. Kumar, "Building segmentation for densely built urban regions using aerial LIDAR data," 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, 2008, pp. 1-8.

[17] K. Kraus and N. Pfeifer. Determination of terrain models in wooded areas with airborne laser scanner data. ISPRS Journal of Photogrammetry and Remote Sensing, 53(4):193-203, 1998.

[18] V. Verma, R. Kumar, and S. Hsu. 3D building detection and modeling from aerial LIDAR data. In CVPR, 2006.

[19] J. Sun, H.-Y. Shum, and N.-N. Zheng. Stereo matching using belief propagation. PAMI, 19:787-800, 2003.

[20] Y. Guo, H. Sawhney, R. Kumar, and S. Hsu. Learning-based building online detection from multiple aerial images. In ECCV, pages 545-552, 2001.

[21] Paolini, L., Grings, F., Sobrino, J., Jiménez Muñoz, J. C., & Karszenbaum, H. (2006). Radiometric correction effects in Landsat multi-date/multi-sensor change detection studies.

International Journal of Remote Sensing, 27(4), 685–704.

[22] Peng, D., Zhang, Y., & Guan, H. (2019). End-to-end change detection for high resolution satellite images using improved UNet++. Remote Sensing, 11(11).

[23] Thome, K., Markham, B., Barker, J., Slater, P. and Biggar, S., 1997, Radiometric calibration of Landsat. Photogrammetric Engineering and Remote Sensing, 63, pp. 853–858.

[24] Zhou, X. (2015). Multiple auto-adapting color balancing for large number of images. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives, 40(7W3), 735–742.

[25] Yao, F., Hu, H., & Wan, Y. (2012). Research on the Improved Image Dodging Algorithm Based on Mask Technique. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XXXIX-B3(September), 519–524.

[26] Tan, K., Zhang, Y., Wang, X., & Chen, Y. (2019). Object-based change detection using multiple classifiers and multi-scale uncertainty analysis. Remote Sensing, 11(3), 1–17.

[27] Zhang, Z., Vosselman, G., Gerke, M., Persello, C., Tuia, D., & Yang, M. Y. (2019). Detecting building changes between airborne laser scanning and photogrammetric data. Remote Sensing, 11(20).

[28] Zhan, T., Gong, M., Liu, J., & Zhang, P. (2018). Iterative feature mapping network for detecting multiple changes in multi-source remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing, 146(January), 38–51.

[29] Khan, S. H., He, X., Porikli, F., & Bennamoun, M. (2017). Forest Change Detection in Incomplete Satellite Images with Deep Neural Networks. IEEE Transactions on Geoscience and Remote Sensing, 55(9), 5407–5423.

[30] Feurer, D., & Vinatier, F. (2018). Joining multi-epoch archival aerial images in a single SfM block allows 3-D change detection with almost exclusively image information. ISPRS Journal of Photogrammetry and Remote Sensing, 146(October), 495–506.

[31] Zhang, X., Shi, W., Lv, Z., & Peng, F. (2019). Land cover change detection from high-resolution remote sensing imagery using multitemporal deep feature collaborative learning and a semi-supervised chan-vese model. Remote Sensing, 11(23).

[32] Low, D. G. (2004). Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 91–110.

[33] Liang, Y., Changjian, W., Fangzhao, L., Yuxing, P., Qin, L., Yuan, Y., & Zhen, H. (2019). TFPN: Twin feature pyramid networks for object detection. Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI, 2019-Novem, 1702–1707.

[34] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2020). Focal Loss for Dense Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2), 318–327.

[35] Zhang, P., Gong, M., Su, L., Liu, J., & Li, Z. (2016). Change detection based on deep feature representation and mapping transformation for multi-spatial-resolution remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing, 116, 24–41.

[36] Fytsilis, A. L., Prokos, A., Koutroumbas, K. D., Michail, D., & Kontoes, C. C. (2016). A methodology for near real-time change detection between Unmanned Aerial Vehicle and wide area satellite images. ISPRS Journal of Photogrammetry and Remote Sensing, 119, 165–186.

[37] Zhang, Y., Gao, J., & Zhou, H. (2020). Breeds Classification with Deep Convolutional Neural Network. ACM International Conference Proceeding Series, 145–151.

[38] Varghese, A., Gubbi, J., Ramaswamy, A., & Balamuralidhar, P. (2019). ChangeNet: A deep learning architecture for visual change detection. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11130 LNCS, 129–145.

[39] Daudt, R. C., Le Saux, B., Boulch, A., & Gousseau, Y. (2019). Multitask learning for large-scale semantic change detection. Computer Vision and Image Understanding, 187, 102783.

[40] Lebedev, M. A., Vizilter, Y. V., Vygolov, O. V., Knyaz, V. A., & Rubis, A. Y. (2018). CHANGE DETECTION IN REMOTE SENSING IMAGES USING CONDITIONAL ADVERSARIAL NETWORKS. International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences, 42(2).

[41] Azar, S. M., Atigh, M. G., & Nickabadi, A. (2018). A multi-stream convolutional neural network framework for group activity recognition. arXiv preprint arXiv:1812.10328.

[42] Chen, H., Wu, C., Du, B., & Zhang, L. (2019, August). Deep siamese multi-scale convolutional network for change detection in multi-temporal VHR images. In 2019 10th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp) (pp. 1-4). IEEE.

[43] Pomente, A., Picchiani, M., & Del Frate, F. (2018, July). Sentinel-2 change detection based on deep features. In IGARSS 2018-2018 IEEE International Geoscience and Remote

Sensing Symposium (pp. 6859-6862). IEEE.

[44] Papadomanolaki, M., Verma, S., Vakalopoulou, M., Gupta, S., & Karantzalos, K. (2019, July). Detecting urban changes with recurrent neural networks from multitemporal Sentinel-2 data. In IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium (pp. 214-217). IEEE.

[45] Lu, D., Mausel, P., Brondizio, E., & Moran, E. (2004). Change detection techniques. International journal of remote sensing, 25(12), 2365-2401.

[46] Rensink, R. A. (2002). Change detection. Annual review of psychology, 53(1), 245-277.

[47] Radke, R. J., Andra, S., Al-Kofahi, O., & Roysam, B. (2005). Image change detection algorithms: a systematic survey. IEEE transactions on image processing, 14(3), 294-307.

[48] Bruzzone, L., & Prieto, D. F. (2000). Automatic analysis of the difference image for unsupervised change detection. IEEE Transactions on Geoscience and Remote sensing, 38(3), 1171-1182.

[49] Chen, G., Hay, G. J., Carvalho, L. M., & Wulder, M. A. (2012). Object-based change detection. International Journal of Remote Sensing, 33(14), 4434-4457.

[50] Wang, Y., Jodoin, P. M., Porikli, F., Konrad, J., Benezeth, Y., & Ishwar, P. (2014). CDnet 2014: An expanded change detection benchmark dataset. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops (pp. 387-394).

[51] Lunetta, R. S., Knight, J. F., Ediriwickrema, J., Lyon, J. G., & Worthy, L. D. (2006). Land-cover change detection using multi-temporal MODIS NDVI data. Remote sensing of environment, 105(2), 142-154.

[52] Rosin, P. (1998, January). Thresholding for change detection. In Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271) (pp. 274-279). IEEE.

２０２０年度　修士論文　深層学習に基づく航空写真とデジタル地表モデルのエンドツーエンドの建物変化検出モデル　　連　欣蕾