END-TO-END BUILDING CHANGE DETECTION MODEL IN AERIAL IMAGERY AND DIGITAL SURFACE MODEL BASED ON NEURAL NETWORKS

Xinlei Lian[1]*, Wei Yuan[1], Zhiling Guo [1], Zekun Cai [1], Xuan Song [1], Ryosuke Shibasaki [1]

[1]The University of Tokyo, Center of Spatial Information Sciences, Japan - (vickie_lxl, miloyw, guozhilingcc, caizekun, songxuan, shiba) @csis.u-tokyo.ac.jp

Commission II, WG II/6

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

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 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. 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.

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). 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.

Within the applications by applying multi-temporal remote sensing imagery to derive timely information on the earth's environment and human activities, most of scholars concentrated on natural environment related 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, urban constructed environment multi-temporal change detection including building construction, traffic construction, urban facilities and other infrastructures timely change is significant for urban activities monitoring, real estate market mastery, resident's mobility and then whole city development promotion. Our study will focus on the application scenario of urban construction change detection including building new construction, demolishment as well as continuation, which is aiming at urban construction legitimacy supervision and real estate commercial activity monitoring.

Along with the developing progression of remote sensing and photogrammetry, land cover change detection is not limited with the dataset utilization on 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 detailed information for spectral 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. 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.

Current 3-D remote sensing-based change detection methods typically appertain to one of following approaches: direct comparison, classification, object-oriented method, model

* Corresponding author

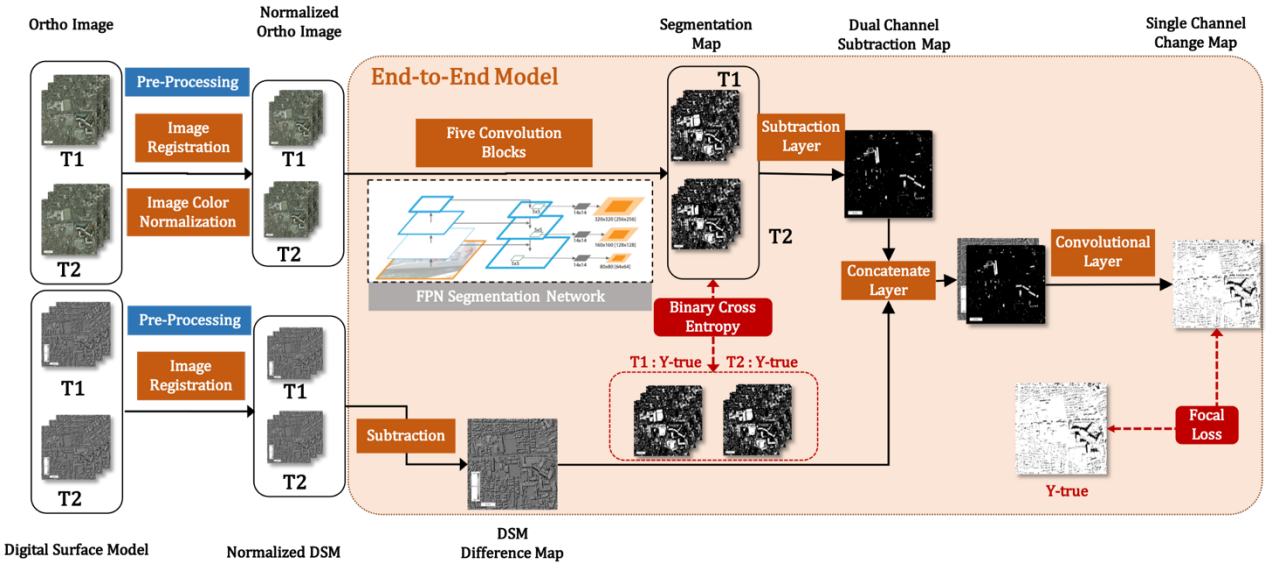

Figure. 1 Overview Framework of Proposed End-to-End Change Detection Model

method and time-series analysis or hybrid method combining two or more of them (Roysam, 2005). With corresponding advantages and shortcomings of great divergence of existing methods, considering the requirements of building change detection research and practical application scenario, 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 of end-to-end network is also appropriate for this task. 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 adjustment space and increase the overall compatibility.

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. 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. The last part of acknowledgement and references will be the final Sections.

2. RELATED WORKS

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).

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). However, in this traditional method, inadequate data quality and fine-tuning of threshold as well as relative parameters remain to be the intrinsic challenge and drawback.

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.

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 accurate performance on the test dataset of natural land-cover imagery with 242spectral 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. Comprehensive change detection model-based researches and methods were innovated but limited to case studies and need to be further explored. End-to-end model for change detection still has enormous development space with divergent 3-D dataset.

3. END-TO-END CHANGE DETECTION MODEL

The overview framework of our proposed end-to-end change detection model is shown in Figure.1. Generally, the proposed framework comprises three components:

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

The model architecture and composition principle will be given in this Section as followed and the algorithm explanation and detailed implementation of the first two components will be given in the next Section of Methodology. 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. The model is modified to accept four inputs including pre-processed ortho image and DSM of T1(Time1) and T2(Time2) and three outputs including building segmentation map of T1 and T2 and the 3-Class Change map.

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.

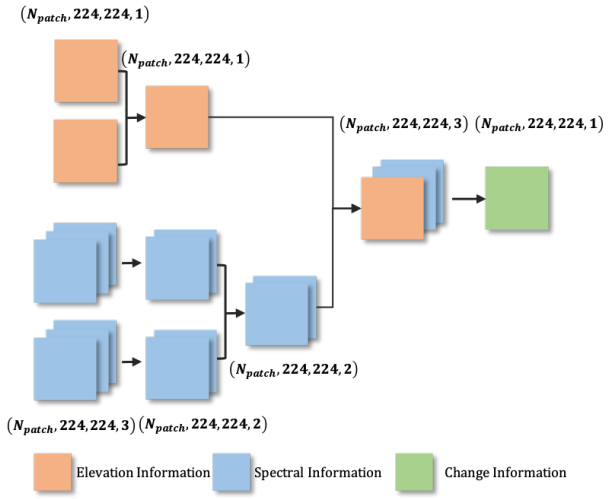

Figure 2. Overall Framework of Shape transformation and feature composition

The framework of shape transformation processing and feature composition is shown in Figure 2. 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 input.

The loss function is another crucial part except model architecture, which comprises two binary cross entropy loss for segmentation intermediate outputs and focal loss for final output as change map, with equal loss weights as [1,1,1]. 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 these circumstances, balancing approaches are brought out. The significance of cross entropy losses is to optimize the segmentation to provide 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 targeted at training data label imbalance problem, which will be introduced in next Section particularly.

4. METHODOLOGY

This section discusses the methodology used to conduct the study. Section 4.1 describes the image-preprocessing measures including image registration and radiometric correction employed in this study. Section 4.2 discusses the characters of FPN as feature extraction network and its superiority for change detection task. Section 4.3 describes the calculation and theory of the novel multi focal loss as a crucial loss function used in model training optimizer.

4.1 Image-Preprocessing

4.1.1 Image Registration

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. Basically, we adapt Scale-Invariant Feature Transform (SIFT) algorithm to corresponding patches of original image for feature key points localization and matching.

The SIFT keypoints are particularly useful for our image registration problem due to their distinctiveness, which enables the correct match for a keypoint to be selected from a large database of other keypoints. This distinctiveness is achieved by assembling a high-dimensional vector representing the image gradients within a local region of the image. The keypoints have been shown to be invariant to image rotation and scale and robust across a substantial range of affine distortion, addition of noise, and change in illumination. Large numbers of keypoints can be extracted from typical images, which leads to robustness in extracting small objects among clutter. The fact that keypoints are detected over a complete range of scales means that small

local features are available for matching small and highly occluded objects, while large keypoints perform well for images subject to noise and blur. Their computation is efficient, so that several thousand keypoints can be extracted from a typical image with near real-time performance on standard PC hardware (David, 2004). Figure 3. shows the key points matching results with ortho aerial imagery of same region in Tokyo taken in 2015 and 2016. The green lines connect corresponding matched key points and a linear homography matrix will be generated for the whole image transformation, aiming at image registration.

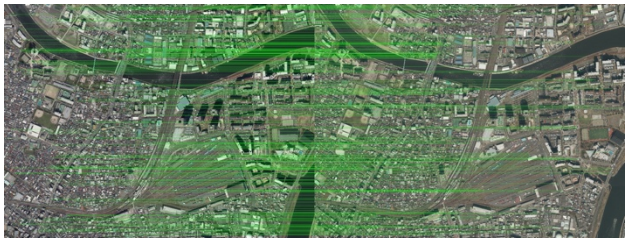

Figure 3. Matched Key points with SIFT on Tokyo Region

4.1.2 Radiometric Correction

Given the fact that the layer parameters will be shared in the feature extraction network for input ortho images of T1 and T2, the radiometric divergence of aerial images caused by disturbing factors will become one of the significant factors during neural network training period. In order to unify the difference in color balance conditions caused by imaging seasons or dates, different solar altitudes and illumination, different angles, different meteorological conditions and different cover areas of cloud, rain or snow etc., for optimized model performance, raising the accuracy and effectiveness, radiometric correction and color normalization methodology will be utilized.

Radiometric corrections serve to remove the effects that alter the spectral characteristics of land features, except for actual changes in ground target, becoming mandatory in multi-sensor, multi-date studies (Paolini, L., 2006). Radiometric correction methods of satellite images can be grouped in two major categories: absolute and relative (Thome et al. 1997). In our study, relative radiometric which is relatively appropriate for multi-temporal radiometric divergence of aerial imagery.

The relative radiometric correction method normalizes images of the same area and different dates by using landscape elements (pixels) whose reflectance values are nearly constant over time. This procedure assumes that the pixels sampled at Time 2 are linearly related to the pixels, of the same locations, sampled at Time 1, and that the spectral reflectance properties of the sampled pixels have not changed during the time interval (no actual change during this period). The sampled pixels are considered pseudo-invariant features (PIFs) and are the key to the image regression method used in the normalization process. (Paolini, L., 2006).

4.2 Feature Pyramid Network (FPN)

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.

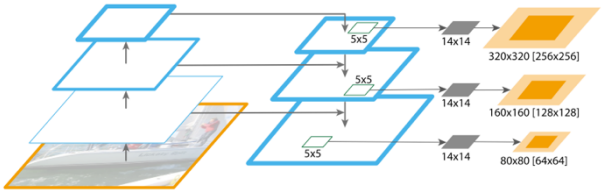

Figure 4. FPN Feature Extraction Architecture for Object Segment

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 4.

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.

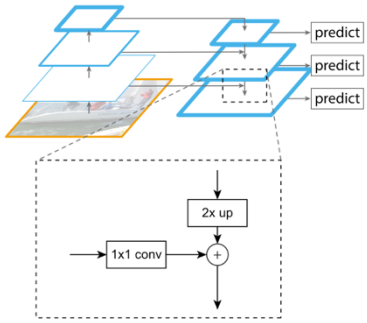

Figure 5. A building block illustrating the lateral connection and the top-down pathway, merged by addition.

For the network architecture, generally, FPN takes a single-scale image of an arbitrary size as input, and outputs proportionally sized feature maps at multiple levels, in a fully convolutional fashion. This process is independent of the backbone convolutional architectures. A bottom-up pathway, a top-down pathway, and lateral connections make construction for FPN pyramid. A building block illustrating the lateral connection and the top-down pathway is displayed in Figure 5. The utilization of FPN acquires the benefits that lateral connections between reconstructed layers and the corresponding feature maps could help the detector to predict the locations more precise, with skipping connections to make training easier.

4.3 Focal Loss for Multi-Class Classification

The Focal Loss (Lin, 2018) is 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:

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

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:

$$p_t = \begin{cases} p & if\ y = 1 \\ 1 - p & otherwise \end{cases} \quad (2)$$

A common method for addressing class imbalance is to introduce a weighting factor α ∈ [0, 1] for class 1 and 1−α for class −1. The α-balanced CE loss is written as:

$$CE(p_t) = -\alpha_t \log(p_t) \quad (3)$$

However, the large class imbalance encountered during training of dense detectors overwhelms the cross-entropy loss. Easily classified negatives comprise the majority of the loss and dominate the gradient. While α balances the importance of positive/negative examples, it does not differentiate between easy/hard examples. Instead, they propose to reshape the loss function to down-weight easy examples and thus focus training on hard negatives. More formally, we propose to add a modulating factor $(1 - p_t)\gamma$ to the cross entropy loss, with tunable focusing parameter $\gamma \geq 0$. In practice, an α-balanced variant of the focal loss is used.

$$FL(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \quad (4)$$

In our end-to-end change detection model, as a multi-class classification problem, an adjusted 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).

$$a = [[a_0], [\alpha_1], [\alpha_2]] \quad (5)$$

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:

$$FLE(\hat{y}, y^t) = \frac{\sum_{i=1}^{N} -(1 - p_i^t)^{\gamma} \log(p_i^t)}{N} \quad (6)$$

where $\hat{y} =$ predicted result tensor
$y^t =$ ground truth tensor
$p^t = \begin{cases} p & if\ y = 1 \\ 1 - p & otherwise \end{cases}$
$N =$ tensor elements number

And the loss calculation we implemented is written as:

$$FLM(\hat{y}, y_{true}) = FLE(\hat{y}, y_{true}) * \alpha \quad (7)$$

5. EXPERIMENTS

Following the principle of inflation study, except for our proposed methodology, two baseline methods are conducted simultaneously for the empirical results, performances of which are also expressed in this Section. 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 difference of dataset and task between our study and other scholars, in order to evaluate the change indicators and validity of procedures, we implemented image differencing post-classification method and similar dual framework with only 2-D imagery, which is following the methodology of Ablation Study.

5.1.1 Post-Classification Method

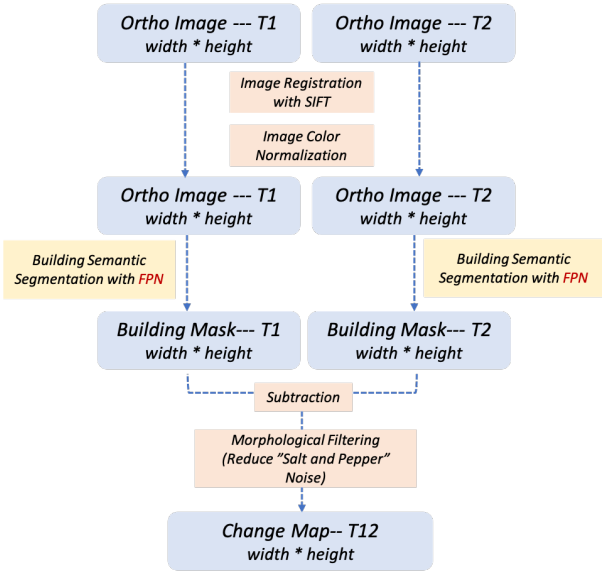

Figure 6. Framework of Post-Classification Methods as Baseline1

Baseline1 could be summarized as a hybrid pixel-based method combining post-classification with segmentation model. Multi-temporal ortho aerial image of T1 and T2, without DSM, will be inputted into FPN semantic segmentation network after pre-processing consisting of SIFT image registration and image radiometric correction. The two outputted building masks will take element-wised subtraction operation. 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. Therefore, the morphological filtering algorithm will be employed to reduce noise for the intermediate result then we will get the final change map. The general framework of baseline1 is shown in Figure 6.

The biggest advantage of this method lies in the employing of existing advanced segmentation network to accurately extract building features. Nevertheless, manual fine-tuning of thresholds and pixel-wised subtraction could not guarantee the robustness of this method.

5.1.2 Dual-Input End-to-End Model

Baseline2 is an end-to-end model-based method with dual inputs and three outputs. 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 and this baseline is mainly for the validation of DSM data introduction effectiveness.

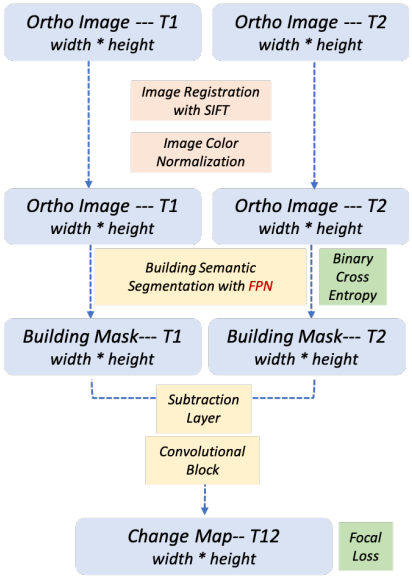

Figure 7. Framework of Dual-Input End-to-End Model Method as Baseline2

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 baseline1 is shown in Figure 7.

5.2 Dataset 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, 12500×9375pixels and 0.16m resolution as well as 1050 raw aerial imagery with sufficient overlap rate over 70%. 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 8. 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.

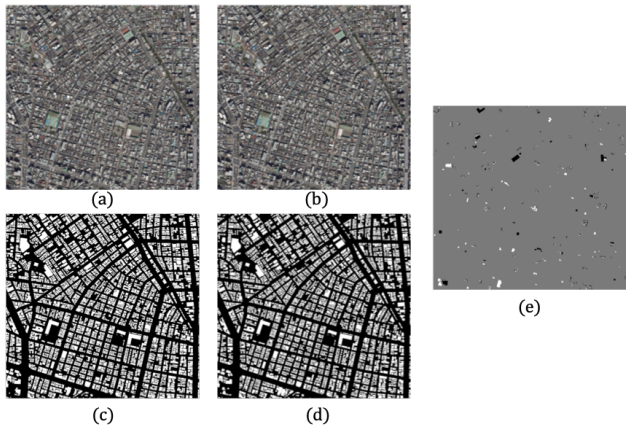

Figure 8. Training Dataset Composition Examples (a) Ortho Image taken in 2015 (b) Ortho Image taken in 2016 (c) Building Segmentation Ground Truth of 2015 (d) Building Segmentation Ground Truth of 2016 (e) Change Map Ground Truth

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

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. For FPN, VGG16 was used as backbone pre-trained model. All convolution operations in the model had a kernel size of 3×3, a stride of 1, and batch normalization. Upsampling operations had kernel sizes including (8,8), (4,4) and (2,2), as well as the nearest-neighbour interpolation method. The input and output tensor dimensionality are (4710, 224, 224, C) where C means channel number differentiates according to data type.

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, open operation was implemented to reduce salt-and-pepper noises with erosion kernel size and dilation kernel size set as 15 and 20 respectively. In baseline2, as an end-to-end model method, it was trained with 2015 and 2016 dataset together, sharing parameters. Our proposed end-to-end model was also trained with 2015 and 2016 dataset including imagery and DSM together. All concatenate operations were implemented by axis 3 and all subtraction layers are merging operation with no parameters trained.

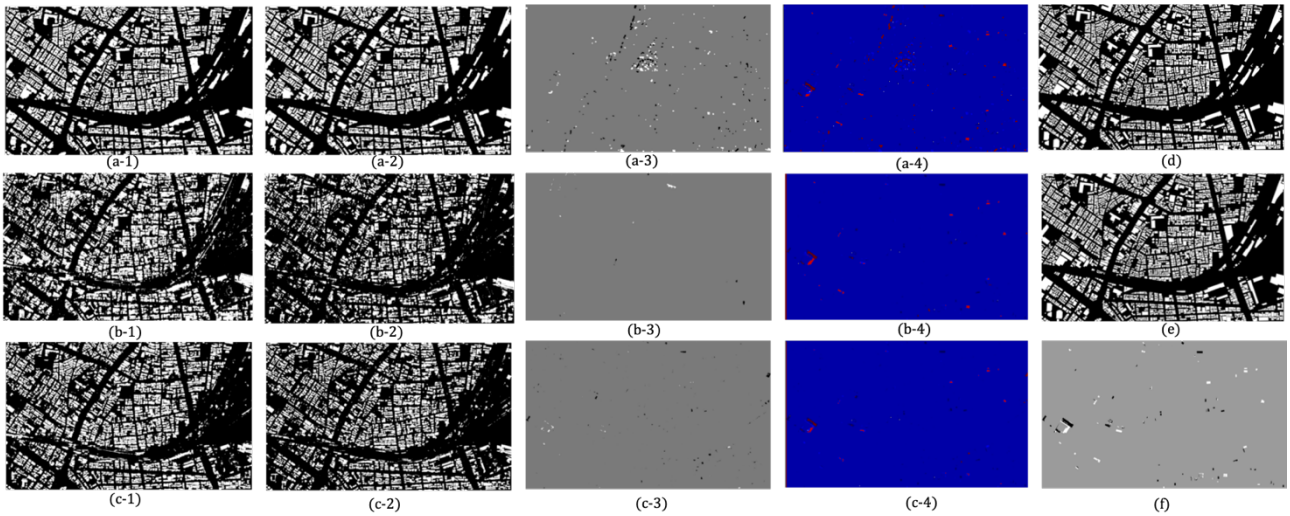

Figure9. Testing Results of Comparison Experiments (a-1) 2015 Segmentation Result of Baselin1 (a-2) 2016 Segmentation Result of Baselin1 (a-3) Change Map Result of Baseline1(a-4) Change Map Visual Evaluation of Baseline2 (b-1) 2015 Segmentation Result of Baselin2 (b-2) 2016 Segmentation Result of Baselin2 (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

Figure 9. 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 Figure1, 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.

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 three evaluation criteria: overall accuracy (OA), precision, recall, F1 score. In their calculation, there are four indexes: 1) true positives, i.e., the number of correctly detected changed pixels; 2) true negatives, i.e., the number of correctly detected unchanged pixels; 3) false positives, i.e., the number of false-alarm pixels; and 4) the false negatives, i.e., the number of missed changed pixels.

The OA is defined as:

$$OA = \frac{TP + TN}{TP + TN + FP + FN}$$

The Precision is defined as:

$$Precision = \frac{TP}{TP + FP}$$

The Recall is defined as:

$$Recall = \frac{TP}{TP + FN}$$

The F1 score is defined as:

$$F1 = 2 * \frac{Recall * Precision}{Recall + Precision}$$

	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

Table 1. Evaluation Metrics of Change Detection Methods

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.

6. CONCLUSIONS AND FUTURE WORKS

This paper proposed an efficient end-to-end model with the utilization of DSM for building change detection task for urban environment, trained and verified on the dataset of Setagaya-Ku, Tokyo aerial imagery. The novel empirical study confirmed that creating a Change Map with proposed method could give relatively high accuracy and performance. End-to-end FPN-based building change detection model offers the following contributions. First, this unsupervised method eliminates costly manual finetuning of thresholds and parameters. The skipping connection structure also offers lower parameter amount and high training effectiveness. Second, adaptive image pre-processing step and weight combination of model optimize the change performance and minimize the noise simultaneously. The future works of this study will focus on more appropriate network layer structure and loss function that is more suitable for change detection task.

REFERENCES

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

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.

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.

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.

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

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.

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.

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.

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.

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.