002

# 2016 年熊本地震後の航空写真を用いた修正 Mask R-CNN モデルに基づく建物抽出と被害評価

# BUILDING EXTRACTION AND DAMAGE ASSESSMENT BASED ON MODIFIED MASK R-CNN MODEL USING POST-EVENT AERIAL IMAGES AFTER THE 2016 KUMAMOTO EARTHQUAKE

19WM1335 戦　義豪
Yihao Zhan

指導教員 丸山　喜久
劉　ウェン

SYNOPSIS

Remote sensing is an effective method for evaluating building damage after a large-scale natural disaster, such as an earthquake or a typhoon. In recent years, with the development of computer vision technology, deep learning algorithms have been used for damage assessment from aerial images. In April 2016, a series of earthquakes hit the Kyushu region, Japan, and caused severe damage in the Kumamoto and Oita Prefectures. Numerous buildings collapsed because of the strong and continuous shaking. In this study, a deep learning model called Mask R-CNN was modified to extract residential buildings and estimate their damage levels from post-event aerial images. Our Mask R-CNN model employs an improved feature pyramid network and online hard example mining. Furthermore, a non-maximum suppression algorithm across multiple classes was also applied to improve prediction. The aerial images captured on April 29, 2016 (two weeks after the main shock) in Mashiki Town, Kumamoto Prefecture, were used as the training and test sets. Compared with the field survey results, our model achieved 95% overall accuracy for building extraction and 88% overall accuracy for the damage classification.

## 1. Introduction

A series of earthquakes hit Kumamoto Prefecture, Japan, in April 2016, including two events: a moment magnitude (Mw) 6.2 foreshock and a Mw 7.0 mainshock. A severe damage to buildings was observed in Kumamoto Prefecture. A total of 8,657 houses completely collapsed, and approximately 190 thousand residential buildings partially collapsed.

It is important to grasp the damage situation immediately after a disaster occurred. Although a field survey could provide more detailed information, it also requires tremendous manpower and time. Under such circumstances, remote sensing technology becomes an alternate way to collect damage information effectively. With the high-resolution aerial images taken by a UAV or aircraft, the image provides a more detailed and richer view of the real world, thereby enabling further analysis of features that are not traditionally visible in satellite imagery.

In the meanwhile, deep learning algorithms as one category of the machine learning methods, have attracted widespread attention in the field of image recognition. Instead of manually operation, deep learning algorithms would learn the image features automatically and output the results.

Combining the benefits of deep learning and remote sensing, our study performs object detection and extraction from wide-area aerial images, and damage classification of residential buildings by using the instance segmentation algorithm Mask R-CNN[1].

## 2. Dataset and Image Processing

Five high-resolution post-event aerial images of Mashiki town were used to create the dataset for building damage assessment. One aerial image consists of 14430 × 9420 pixels. The aerial images were taken by the Geospatial Information Authority of Japan (GSI) using UltraCamX on April 29, 2016, two weeks after a series of earthquakes. The five aerial images were mosaiced and covered the center of Mashiki town, as shown in **Figure 1**. The most affected area was selected in the training and test sets. In **Figure 1**, the training area is covered with red, whereas the green color represents the test area.

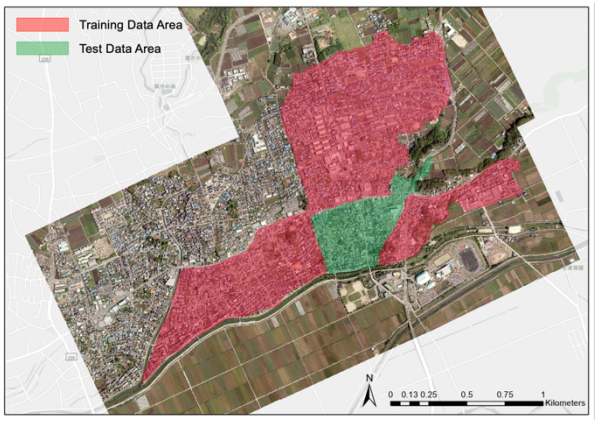


**Figure 1** Training and test sets generated from the five high-solution aerial images taken by GSI on April 29, 2016

Unlike the dataset of natural images, the viewpoint of remote sensing image datasets usually rests on the top, which makes the

target objects appear relatively small. A single large-scale aerial image can contain over thousands of buildings. In this case, we cut the original image into 500 × 500 pixel square images to obtain a total of 510 training images and 157 test images. The buildings located at the edge of the cut images would appear in two or more images. In this case, the key feature might be separated at different images, which leads to a decrease in detection accuracy. To solve this problem, we mark their complete shapes by shifting the cutting frame.

These cut images were labelled manually by the LabelMe tool into four damage categories. The building damage categories were cited using the resources of the Architectural Institute of Japan (AIJ), from the report of the Ministry of Land, Infrastructure, Transport and Tourism (MLIT)[2]. Kumamoto Earthquake Disaster Investigation Committee of the Kyushu Branch of the AIJ surveyed 2,652 buildings and classified them with damage grade based on the research of Okada and Takai[3]. A detailed description of the damage classifications is shown in **Table 1**. Several samples of the labelled buildings are shown in **Figure 2**.

**Table 1.** Definition of the damage grades used in this study

| Damage grades | Damage grades (MLIT[2]) | Okada, Takai[3] |
|---|---|---|
| Level_1 | No damage | D0 |
| Level_2 | Slight damage | D1 – D3 |
| Level_3 | Severe damage | D4 |
| Level_4 | Collapsed | D5 – D6 |


**Figure 2**. Label examples of the buildings in levels 1 to 4

## 3. Model and Modification

The workflow of Mask R-CNN is as follows: 1) Mask R-CNN feeds the image to the residual network to extract features and generate multiscale feature maps; 2) side-joining is performed, and the feature maps at each stage are upsampled twice and tensor-summed with the adjacent underlying layers; 3) the feature maps are fed into RPN to generate candidate regions on the feature maps with different sizes that are input along with the feature maps to RoI Align to obtain the bounding boxes; and 4) the bounding boxes are classified and regressed, and a high-quality instance segmentation mask of the detected object is generated. The structure shown in **Figure 3**.

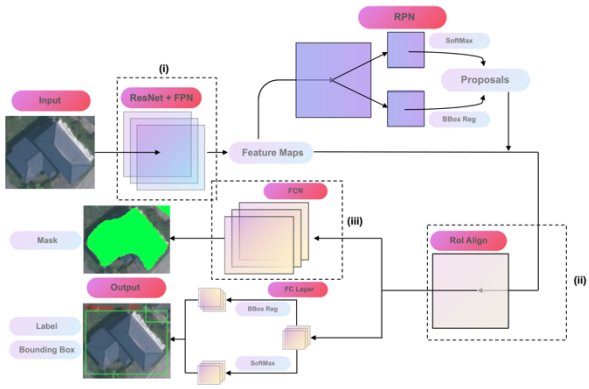


**Figure 3**. Structure of the Mask R-CNN algorithm

Although Mask R-CNN has a high-level capability in object detection, it still has some shortcomings when applied directly to damage detection from aerial images:

1. The localization information for large objects is sometimes inaccurate, while the details of medium and small objects may sometimes be lost in the feature map.
2. The target buildings that cover a small proportion of the whole image will generate more negative samples in the region proposal layer, which will decrease the accuracy of the model.
3. The two different outputs may overlap with each other for the same object.

To get better detection performance of Mask RCNN algorithm in aerial images, we employed some modification on the original network shown in **Figure 3**(i), (ii) and (iii). The first modification in **Figure 3**(i) is Path Aggregation Network[4]. The basic idea of this modification is to shorten the information transmission path and use the precise location information of the lower-level features fully by adding bottom-up branches with reverse lateral connections to the feature pyramids. The second modification in **Figure 3**(ii) is Online Hard Example Mining[5]. This method solves the problem of imbalance between positive and negative samples during training by expand the original RoI network into two RoIs, one RoI with only forward propagation for calculating the loss and one RoI with normal forward-backward propagation, using the hard example as the input to calculate the loss and pass the gradient. The third modification is called multiclass Non-Maximum Suppression. It is a piece of code that we wrote ourselves to fix the overlap of the result. We wrote an Intersection over Union (IoU) checking process in **Figure 3**(iii). Only when the IoU of two bounding box exceeds a threshold value, the bounding box with the low confidence score will be removed.

## 4. Evaluation and Discussion

The latest object detection works tend to use the COCO dataset to demonstrate the effectiveness of their models. For the COCO dataset, an interpolated AP calculation is used by sampling 100 points on the PR curve. Moreover, the threshold of IoU is taken in the range of 0.5-0.95 with intervals of 0.5, and average of AP values are calculated with these settings [52]. This average AP (mean AP) value will be taken as the final result.

Due to the narrow spacing between the buildings and unregular shapes, before training our modified model, we ran several tests on the original Mask R-CNN model to determine

the best RPN parameters for this dataset. The mAP of model-1 with adapted RPN parameters is approximately 4% higher than that of the original model. Then, we experimented with several modified models based on the Mask R-CNN model-1. Several combinations of PANet and OHEM with different epochs are tested, for which results are shown in **Table 2**.

**Table 2** Comparison of the test results using modified models based on the Mask R-CNN model-5

| Model | Epochs | PANet | OHEM | Bounding Box mAP | Segment mAP |
|---|---|---|---|---|---|
| 1 | 300 | No | No | 0.332 | 0.333 |
| 2 | 80 | No | No | 0.345 | 0.342 |
| 3 | 80 | No | Yes | 0.350 | 0.356 |
| 4 | 80 | Yes | No | 0.352 | 0.368 |
| 5 | 80 | Yes | Yes | 0.361 | 0.370 |
| 6 | 40 | Yes | Yes | 0.365 | 0.373 |

When PANet and OHEM were applied independently, they both improved the results by approximately 2%. When they were both applied, the improvement reached 3-4%. Finally, we performed additional tests on model-5 with different numbers of epochs and chose 40 epochs as the final epoch value. The mAP of the bounding box was 0.365 and that of the segmentation was 0.373 in model-6.

In the test area, 95.1% of the buildings were identified successfully. The precision of building detection was 91.4%. Even the severely damaged buildings (Level_3 and Level_4) could be extracted with 94.3% accuracy. The precision of building detection reached 92.0%. The misdetections were caused by delineations between two buildings or hidden shadows of other buildings. On the other hand, overdetection was caused by buildings that were not investigated by the AIJ. However, our model detected those buildings and estimated their damage grades. The classification precision of Level_1 exceeded 83%, and recall was 76%. The precision and recall of Level_2 were 72% and 88%, respectively. The precision of Level_3 was 83%, and recall was greater than 70%. For the total collapsed buildings in Level_4, the precision exceeded 93%, and recall was approximately 85%. The OAA of the classification reached 82%. The confusion matrix for the test area is shown in **Table 3**, which showed acceptable results.

**Table 3**. Confusion matrix of the damage classification for the buildings in the test area

| | | **Prediction for the test area** | | | |
|---|---|---|---|---|---|
| | | Level_1 | Level_2 | Level_3 | Level_4 |
| **True Label** | Level_1 | 34 | 11 | 0 | 0 |
| | Level_2 | 7 | 71 | 1 | 2 |
| | Level_3 | 0 | 9 | 28 | 3 |
| | Level_4 | 0 | 8 | 5 | 75 |

Based on these results, we applied model-10 to the whole target area of Mashiki town and obtained a prediction map of building damage, as shown in **Figure 4**. From the prediction map, we can see that most of the collapsed buildings (Level_4) were around the No. 28 Prefectural Road. More than half of the buildings were classified into Level_1 and Level_2, indicating no damage or less than moderate damage, respectively.

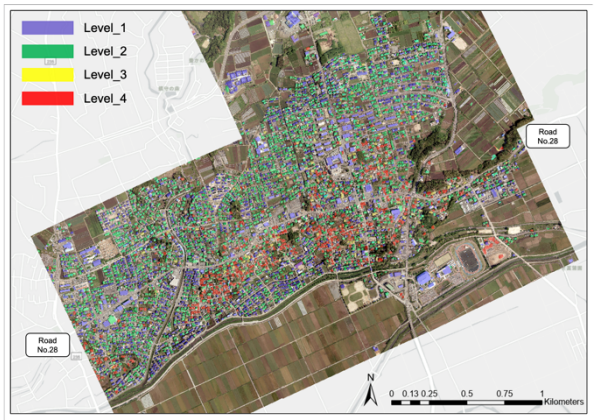


**Figure 4**. Obtained prediction map of all the buildings in the target area using the proposed model-6.

To verify the accuracy of our prediction map, we compare it with the report of the AJI. In the 2016 Kumamoto Earthquake Report, the ratios of collapsed buildings were summarized in 57-m grids. The ratio of the number of collapsed buildings to that of all buildings in the grid was calculated. The grid map of the collapsed ratio is shown in **Figure 5(a)**. There were 414 grids in the target area of the field survey, which were classified into five classes according to the collapse rates. There were 262 grids in the class of 0%, 47 grids in the class of 0-25%, 53 grids in the class of 25-50%, 37 grids in the class of 50-75% and 10 grids in the class of 75-100%. To verify our predication results, we calculated the ratio of Level_4 buildings in **Figure 4** using the same grid scale, and it is shown in **Figure 5(b)**. There are 260 grids in the class of 0%, 45 grids in the class of 0-25%, 62 grids in the class of 25-50%, 36 grids in the class of 50-75% and 11 grids in the class of 75-100%.

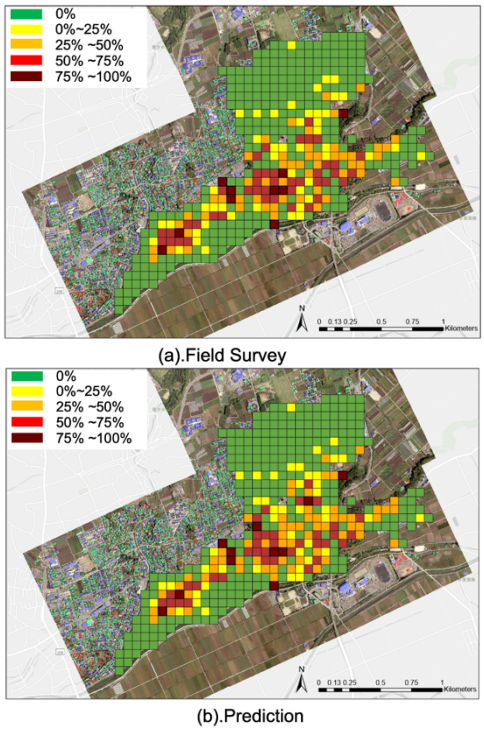


**Figure 5**. Comparison of (a) the collapsed ratio in the report of the field survey and (b) our prediction in a 57-m grid unit.

For each class of collapsed ratios, we obtained a decent result. The precision and recall for the collapsed ratio of 0% reach 99.2% and 98.9%, respectively. For the collapse ratio of 0-25%, the precision is 93.3%, and the recall is 89.4%. Additionally, the grids associated with the collapse ratio of 25-50% have 85.3% and 98.1% precision and recall, respectively. The precision and recall for a collapse ratio of 50-75% reach 100% and 88.5%, respectively. For the collapse ratio of 75-100%, the precision is 81.8%, and the recall is 90.0%. The OAA of all the classes of collapsed ratios reached approximately 96%. The confusion matrix is shown in **Table 4**. Most of the grids were classified either in the accurate class or neighbouring classes. A single non-collapsed grid was mistakenly classified as 25%−50%, and another grid with a collapse ratio of 75%−100% was underestimated as 25%−50%. These were both caused by the differing counts of buildings. While some houses in Japan have primary buildings, secondary buildings, and warehouses, the field survey from the AIJ only recorded damage to primary buildings. In addition, unoccupied houses were not counted in the report. This led to the difference between the prediction results and the true data.

**Table 4**. Confusion matrix of grid prediction in the whole investigated area

| | | Prediction | | | | |
|---|---|---|---|---|---|---|
| | | 0% | 0%−25% | 25%−50% | 50%−75% | 75%−100% |
| **Field Survey** | 0% | 259 | 2 | 1 | 0 | 0 |
| | 0% − 25% | 2 | 42 | 3 | 0 | 0 |
| | 25% − 50% | 0 | 1 | 52 | 0 | 0 |
| | 50% − 75% | 0 | 0 | 4 | 37 | 1 |
| | 75% − 100% | 0 | 0 | 1 | 0 | 9 |

Since the prediction map in **Figure 5** included the area used for the training set, we selected the grids containing only the test area. In a total of 37 grids reported by the field survey, there were 9 grids with the collapsed ratio of 0%, 7 grids with that of 0-25%, 9 grids with that of 25-50%, 9 grids with that of 50-75%, and 3 grids with that of 75-100%. For our prediction results, 9 grids were classified as the collapsed ratio of 0%, 7 grids as 0-25%, 12 grids as 25-50%, 7 grids as 50-75% and 2 grids as 75-100%. The precision and recall for the grids with collapse ratios of 0% and 0-25% were 100%. The precision and recall for the grids with collapse ratios of 25-50% were 75%, and the recall remained at 100%. For the grids with collapse ratios of 50-75% and 75-100%, the precision reached both 100%, and the recall reached 77.78% and 66.67%, respectively. The OAA of prediction for the test area reached 91%, which showed a high capability to detect severely damaged areas.

We also compared our results with previous studies for damage estimation of the buildings after the 2016 Kumamoto earthquake. Liu et al. achieved an OAA of approximately 46.8% in the classification of damaged buildings by using multitemporal PALSAR-2 data[6]. Although SAR images can be obtained despite weather conditions, damage assessment from SAR images with high accuracy is still difficult. Naito et al. developed a CNN model that performed an 88.4% OAA for the damage classification task based on the evaluation indicators of personal visual interpretation[7]. Our model achieves the same level of OAA as 88.1% by using the report of the MLIT field survey. Compared with the verification using visual interpretation, the performance of our model is more objectively evaluated. In addition, our modified NMS approach gave a prediction map with an accuracy of 96% OAA for the collapsed ratio, which would be useful for the emergency response after an earthquake.

## 5. Conclusions

In this study, we aim to develop a faster means to extract damaged buildings and estimate their damage levels using high resolution aerial images taken after the 2016 Kumamoto earthquake. We modified the original Mask R-CNN model by adding PANet, OHEM and the modified NMS, which improved the mAP of the model from 29% to 37% and enhanced the capability of detecting small objects with similar features. By training and testing the aerial images, 95.1% of the buildings were detected successfully, and the overall accuracy of the damage classification was 88%.

The best model was applied to the entire target area of Mashiki Town, Kumamoto Prefecture, Japan, and a prediction map of building damage was obtained. The prediction results were verified by comparison with the field survey report of the MLIT. The ratio of the collapsed buildings in the 57-m grids was calculated. The overall accuracy of the whole grid was approximately 96%, whereas the accuracy for the grid containing only the test area was approximately 91%. The results support that the proposed model is a viable method of quickly extracting damaged buildings due to earthquakes from postevent aerial images. For further research, we will focus on different regions or different natural disasters by using transfer learning to create a general solution for the damage detection after natural disasters.

## REFERENCES


1) He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
2) Ministry of Land, Infrastructure, Transport and Tourism (MLIT). Report of the Committee to Analyze the Causes of Building Damage in the Kumamoto Earthquake (In Japanese). 2016. pp27-38.
3) Okada, S.; Takai, N. Classifications of Structural Types and Damage Patterns of Buildings for Earthquake Field Investigation. Journal of Structural and Construction Engineering (Transactions of MLIT). 1999, 64 (524), 65-72.
4) Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8759-8768.
5) Shrivastava A.; Gupta, A.; Girshick, R. Training Region-Based Object Detectors with Online Hard Example Mining. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 761-769
6) Liu, W.; Yamazaki, F. Extraction of collapsed buildings in the 2016 Kumamoto earthquake using multi-temporal PALSAR-2 da-ta. Journal of Disaster Research. 2017, 12, 241-250.
7) Naito S.; Tomozawa, H.; Mori, Y.; Nagata, T.; Monma N.; Nakamura, H.; Fujiwara, H.; Shoji, G. Building-damage Detection Method Based on Machine Learning Utilizing Aerial Photographs of The Kumamoto Earthquake. Earthquake Spectra. 2020, 36(3),1166-1187.