DETECTING CRACKS AND SPALLING AUTOMATICALLY IN EXTREME EVENTS BY END-TO-END DEEP LEARNING FRAMEWORKS

Yongsheng Bai[1*], Halil Sezen[1], Alper Yilmaz[1]

[1] Dept. of Civil, Environmental and Geodetic Engineering, The Ohio State University, 2070 Neil Avenue, Columbus, Ohio, USA
–(bai.426, sezen.1, yilmaz.15)@osu.edu

COMMISSION II, WG II/5

KEY WORDS: Deep Learning, Mask R-CNN, Cracks Detection, Spalling Detection, Structural Damage Detection

ABSTRACT:

In this paper, we develop and implement end-to-end deep learning approaches to automatically detect two important types of structural failures, cracks and spalling, of buildings and bridges in extreme events such as major earthquakes. A total of 2,229 images were annotated, and are used to train and validate three newly developed Mask Regional Convolutional Neural Networks (Mask R-CNNs). In addition, three sets of public images for different disasters were used to test the accuracy of these models. For detecting and marking these two types of structural failures, one of proposed methods can achieve an accuracy of 67.6% and 81.1%, respectively, on low- and high-resolution images collected from field investigations. The results demonstrate that it is feasible to use the proposed end-to-end method for automatically locating and segmenting the damage using 2D images which can help human experts in cases of disasters.

1. INTRODUCTION

Automation on Structural Damage Detection (SDD) and Structural Health Monitoring (SHM) is made possible with the rapid development of vision- and vibration-based technologies. The necessity of using them to assist human experts is to increase the accuracy, rapidness and efficiency of SSD and SHM while reducing the overall cost. With the successful application of deep learning methods on a wide range of problems, it is imperative to apply these techniques on SDD and SHM. Generally speaking, the application of deep learning on SDD and SHM requires an interdisciplinary team. These teams typically use low-cost sensors and autonomous platforms such as Unmanned Aerial Vehicles (UAVs) and Unmanned Ground Vehicles (UGVs) in field inspections for real-time inspection and monitoring.

Detection and identification of structural damage can typically be performed by image segmentation and image classification. In case of classification, the goal is to identify the categories of structural attributes, such as material types (e.g., steel, concrete, masonry) or structural damage types (e.g., cracks, spalling, collapse) without locating the position of damage from images. On the other hand, the goal of image segmentation is to detect and mark damage in specific regions where each pixel in the image is labeled to denote types of material failures, such as cracks, spalling and other indicators of structural failures. Spalling refers to the concrete cover of the steel reinforcements or part of nonstructural and structural materials that was split and separated from the original materials. Cracks, on the other hand, are the phenomena of discontinuity of materials observed on the surface of them.

Structural damage can appear in images in different ways and at various scales. Damage can span a larger or smaller extent, or even be invisible (Gao and Mosalam, 2018). In addition, the image resolution may also cause problem for the same type of damage when the deep learning models are utilized to detect it (Bai et al., 2020b). Therefore, it is necessary to develop a robust end-to-end solution to segment structural damage automatically. Mask R-CNN has recently been successfully applied to instance segmentation in computer vision (Cai and Vasconcelos, 2019). Based on its success, this approach is adopted to segment damage so that the buildings and bridges can be continuously monitored. In particular, three variations of Mask R-CNN networks are proposed to detect two major types of structural damage, spalling and cracks, that works independent of scale and image resolution. Publicly available image datasets collected from field investigations in recent large earthquakes are used to check the effectiveness of the models.

2. RELATED WORK

2.1 Deep learning with R-CNNs in image segmentation

There are several major deep learning methods for image segmentation, including but not limited to Fully Convolutional Networks (FCNs), encoder-decoder models, multi-scale and pyramid networks, Regional Convolutional Neural Networks (R-CNNs), etc. Each approach has its own advantages, and some are typically used in benchmarking studies (Minaee et al., 2020).

Multiple convolutional layers are typically utilized as feature extractors while downsampling and then upsampling the data within sliding windows. Its efficiency has been shown to be low when FCNs are used. R-CNNs, on the other hand, can preprocess the input image to produce thousands of Region of Interests (RoIs) for feature extraction with FCNs. Furthermore, the R-CNN reduces the computational time compared to alternative approaches and improves the accuracy of segmentation. Its computational cost, however, is still high. To improve it, Fast R-CNN and Faster R-CNN have been

[*] Corresponding author

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-2-2021-161-2021 | © Author(s) 2021. CC BY 4.0 License.

introduced and their structures are quite different from the conventional R-CNN (Ren et al., 2015). The former applies FCNs directly on the RoIs of the feature maps which comes after convolutional process on the original image. A network referred to as Region Proposal Network (RPN) on the feature maps is inserted to automatically produce the region proposal in the case of Faster R-CNN. Thus, it improves the speed and accuracy of prediction. But neither of these solutions are applicable to instance segmentation. He et al. (2017) proposed a benchmark network, Mask R-CNN, to predict the instance as well as its bounding box and class. Recently, several variations of Mask R-CNN were published where researchers use different backbone network architectures for feature extraction, some of which we adopted and developed for detecting spalling and cracks automatically in this paper.

2.2 Spalling and cracks detection with deep learning

Several researchers have adopted deep learning methods for detecting structural damage. Hoskere et al. (2018) illustrate an experiment with 23-layer ResNet and 9-layer VGG networks to classify and segment seven classes of structural damage, including cracks, spalling, exposed reinforcement, corrosion, fatigue cracks, asphalt cracks, and no damage. Ali et al. (2019) introduce Faster R-CNN for defect detection in historical masonry buildings with high resolution images. Kong and Li (2018) describe an application that detects and tracks the propagation of cracks in a steel girder in image streams. Atha et al. (2018) explain the difference between two CNN methods used in detecting metallic corrosion. Gao and Mosalam (2020) started the Phi-Net Challenge for collecting pictures of building structural failures in 2018. Their large dataset, which is also used in this paper, is suitable for training and testing different methods for structural damage detection at different scales (Bai et al., 2020b).

Recent research in image segmentation have significantly advanced application of deep learning on structural damage detection. Yang et al. (2018) employed a hybrid network, composed of Holistically-Nested Edge Detection (HED) network and U-Net to detect cracks and spalling on concrete structures, and then reconstruct 3D model through Simultaneous Localization and Mapping (SLAM) for UAV images. Cha et al. (2018) applied Fast R-CNN on detecting five types of structural damage, including concrete cracks, steel corrosion of two levels (medium and high), bolt corrosion, and steel delamination. For this purpose, authors labelled 2,366 images with the size of 500×375 for training. Attard et al. (2019) trained a Mask R-CNN with 200 images to locate cracks on the concrete surface at pixel level. Kim and Cho (2019) used 376 images in their training data for Mask R-CNN to find the cracks on a concrete wall with high resolution cameras and utilized an additional image processing procedure on each bounding box to quantitatively measure the width of these cracks. Kalfarisi et al. (2020) introduced structured random forest edge detection into bounding boxes of a Faster R-CNN to detect cracks on infrastructures and compared it with the performance of Mask R-CNN. A total of 1,250 images were included in training and validation process with the size varying from 344×296 to 1,024×796. These models are verified with images acquired from field inspections on structural members, including building walls, bridge columns, tunnel walls and roads. The results show that both approaches are robust for this task. Finally, they used photogrammetry software to construct a 3D reality mesh model so that the cracks can be visualized and quantified further. Mondal et al. (2020) used Faster R-CNN to automatically detect four common types of structural damage, including surface cracks, spalling (which includes facade spalling and concrete spalling), and severe damage with exposed rebars and severely buckled rebars, but they didn't mark the enclosing regions of these damage. Instead, they used bounding boxes to give the scope of them.

Based on conclusions drawn from the aforementioned papers, some researchers also have started to conduct their studies on defect detection, identification and localization. Some cited researchers prefer to only classify the structural damage as it does not require time-commitment for labeling the damage (Zha et al., 2019). As a result, the location and position of cracks on structural components or structures are unknown until a human expert manually checks and marks them out. Simplicity of annotation process makes these models to be trained with a large number of images which is not the case for segmentation networks that faces problems due to insufficient training samples. In order to inherit the advantage of classification networks, it is necessary to employ a segmentation network to locate the damage once various types of structural damage have been classified.

In a recent study, a cascaded network that includes a ResNet and a U-Net to detect cracks (Bai et al., 2020b) is proposed. A 152-layer ResNet which meets the accuracy requirements of identifying scene level, material types, and damage types, is applied at the first step. Even the severity of structural damage can be quantified by it (Zha et al., 2019). U-Net has been utilized in the second step to mark the damage region, such as cracks, at various image scales. Tests on public datasets have shown that the cascaded network improves the accuracy of the detection dramatically in larger scale detection tasks (Bai et al., 2020b). The Cascaded network, however, take a long time to detect the cracks. Therefore, two end-to-end networks are introduced, such as one of Mask R-CNN with attention mechanism and Path Aggregation Network (PANet), and the other with a new backbone called High-resolution Network (HRNet) (Bai et al., 2020a). Tests for crack detection have shown that these new models can achieve an accuracy of 75.1% with 2,021 labeled images for training and validation. Both of the networks are used as the primary methods in this paper. Moreover, another method named Cascade Mask R-CNN (Cai and Vasconcelos, 2019) is also employed here. New training and validation images are curated for spalling and cracks detection.

3. METHODOLOGY

In this section, the dataset and the network structures we utilized are introduced along with the final architecture used to solve the problem at hand. These methods are chosen here because two of them have a good performance in our previous study (Bai et al., 2020).

3.1 Data preparation and augmentation

In training process, a dataset similar to Common Objects in Context (COCO) is generated from the public sites and from Yang et al. (2018). The images in this dataset are labeled by the tool referred as the COCO Annotator (Brooks, 2019), in which the polygons are used to define the boundaries of the cracks and spalling and the closed region of these polygons are the damage in images. Some examples from this process are shown in Figure 1. In these labeled images, cracks, spalling and background are in yellow, green and purple, respectively. Size of the training and labeling images varies from 147×288 to 4600×3070. By excluding steel structures, these surface cracks on structural or nonstructural materials are at various scales, and the reinforcements may be exposed or not in spalling cases.

In order to increase the training dataset, albumentation is employed for data augmentation. Buslaev et al. (2020) develop this method with pixel-level transformation and spatial transformation, including flipping, rotating, cropping, etc. Spatial transformation is adopted to preprocess our training data since it can change the input images, masks and bounding boxes simultaneously.

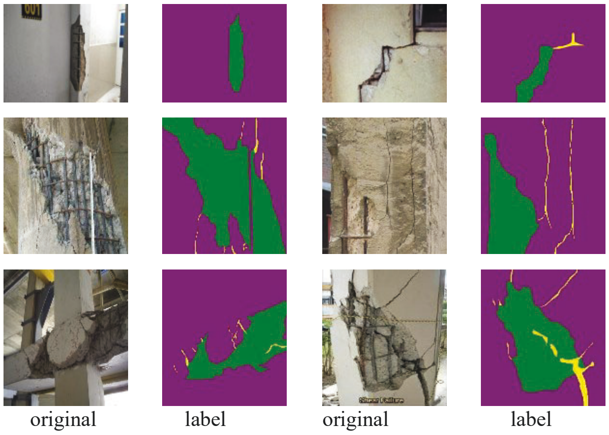


Figure 1. Some examples of training data.

3.2 Mask R-CNN with Path Aggregation Network (PANet) and Spatial Attention Mechanisms

He et al. (2017) proposed Mask R-CNN for instance segmentation, which is an extension of Faster R-CNN. A RPN is inserted onto feature maps to automatically produce RoIs, then a small FCN is applied on each RoI to segment the instance of objects with masks. In addition, different depth of ResNet and the Feature Pyramid Network (FPN) are combined to extract high-quality feature maps. Since Mask R-CNN is a benchmark for instance segmentation in image processing, many improvements have been made since its publication. The framework of Mask R-CNN is shown in Figure 2. Liu et al. (2018) improved Mask R-CNN by replacing FPN with PANet to improve performance. Because features of low layers in the pyramid can reach high layers by skip-connections and a technique called adaptive feature pooling can fuse all levels of features for each proposal, their proposed method achieves a higher accuracy when a modified approach on mask prediction is adjusted. Figure 3 shows the framework of PANet, which is in part used in our paper. Furthermore, we also introduce spatial attention mechanisms as suggested by Zhu et al. (2019) into our approach. The goal of this study is to facilitate the backbone of Mask R-CNN to extract more useful features in cracks and spalling detection. This method is called as APANet Mask R-CNN in this paper.

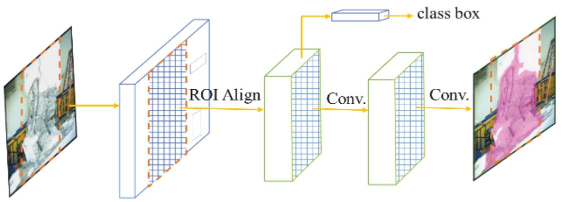


Figure 2. The Mask R-CNN framework for instance segmentation of structural damage.

3.3 Mask R-CNN with High-resolution Network

CNN can have a number of different backbones when applied for segmentation problem. For example, the original Mask RCNN uses a 101-layer ResNet as its backbone. But Sun et al. (2019) developed a new network named HRNet to extract features from an original image. Utilizing repeated multiscale fusions across these convolutional blocks, this network maintains high-resolution representations via inter-connections between high- and low-resolution convolutional modules within a parallel structure. As shown in Figure 4, there are four stages in HRNet. High-resolution features are kept until the end of convolutional operation, and low-resolution ones are added to each new stage. The connection between them may be the key for better feature extraction. In this paper, HRNet is employed as the backbone of another Mask R-CNN to detect the aforementioned two types of structural failures, spalling and cracks. This approach is named as HRNet Mask R-CNN for this study.

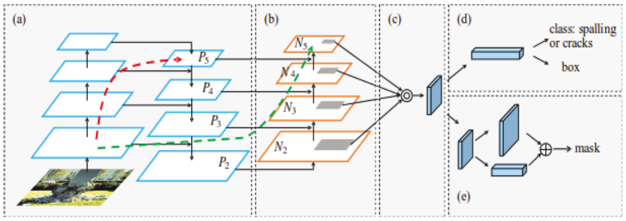


Figure 3. Illustration of the framework of PANet for SDD. (a) FPN backbone. (b) Bottom-up path augmentation. (c) Adaptive feature pooling. (d) Box branch. (e) Fully-connected fusion. *Pi* and *Ni* are *ith* of the original pyramid layers and new feature layers, respectively.

3.4 Cascade Mask R-CNN

Cascade Mask R-CNN solves the overfitting problem when a larger threshold used to compute Intersection of Union (IoU) and disproportion of the quality over the inference and training when Mask R-CNNs are used (Cai and Vasconcelos, 2019). The stages of object detection architecture are increased from two to four on processing object proposals after features are extracted by CNN from the original input image. For a typical Mask R-CNN shown in Figure 5, H0 is the feature proposal network to produce massive proposals for each ROI and H1 is the RPN for automatically generating accurate candidate proposals. B, C and S respectively denote bounding box, class score and segmentation branch. Cascade Mask R-CNN increases the stages to combine these candidate bounding boxes in previous stage and features are resampled from the feature map in the next step. There are different strategies to insert the regression process on segmenting the instances, but the final mask prediction is the result from the single segmentation branch of Figure 5(b) and 5(c), and from three segmentation branches of Figure 5(d). Thus, position of bounding boxes and class scores can be kept consistent and the regions of the instances can be refined to be more accurate. This method focuses on improving the capability of the detector to find better candidate bounding boxes, class scores and mask predictions.

4. IMPLEMENTATION

In our study, we adopt the source codes of the Mask R-CNNs provided in MMDetection (Chen et al., 2019). There are some modifications including revising part of the program and finetuning parameters during the training and testing. The augmented images are provided to the modified models to train

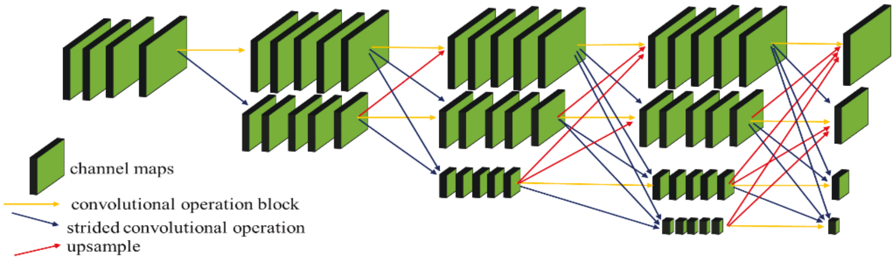

channel maps
convolutional operation block
strided convolutional operation
upsample

Figure 4. Framework of high-resolution network (HRNet). There are four stages. The 1st stage consists of high-resolution convolutions. The 2nd (3rd, 4th) stage repeats two-resolution (three-resolution, four-resolution).

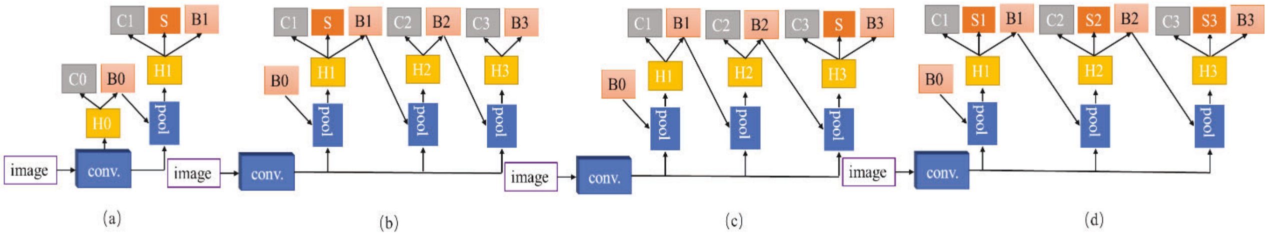


Figure 5. Mask R-CNN (a) vs. three Cascade Mask R-CNN strategies for instance segmentation (b)-(d). "I" is input image, "conv" is backbone convolutions, "pool" is region-wise feature extraction, "H" is network head, "B" is bounding box, "C" is classification, and "S" denotes a segmentation branch. Note that the segmentation branches do not necessarily share heads with the detection branch.

and evaluate at first. The data from Phi-Net dataset (Gao and Mosalam, 2020), 2017 Pohang earthquake dataset (Sim et al., 2018) and 2017 Mexico City earthquake dataset (Purdue University, 2018) at various scales are used to test the algorithm. In the training process, the hyperparameters are set as: learning rate is 0.002, momentum is 0.9 and decay rate of weights is 0.0001. The loss function for mask is cross-entropy and for bounding boxes is smooth L1. The training and testing for the model are executed with NVIDIA GeForce GTX 2080 Super. Total number of epochs for training each model is set to 100. Based on testing on our own data, all above parameters are finally selected after we compared and optimized them.

4.1 Evaluation on the proposed models

Considering that the labeled data of our training and validation are similar to COCO dataset, we follow the same standard metrics to evaluate our models based on our validation dataset. The results can be show in Table 1, AP (Average Precision) is based on IoU, different threshold values and various scales provide [AP_{50}, AP_{75}, *AP*, AP_S, AP_M, AP_L].

Mask AP (box) is reported here. Both Cascade Mask R-CNN and APANet Mask R-CNN employ 101-layer ResNet as backbone. HRNet Mask R-CNN uses four-stage high resolution networks. But the *APs* of these Mask R-CNNs on the validation data are very low (see Table 1 and 2), even though they are close. For damage detection, it is more important to identify and mark the damage as many and precisely as possible for large image datasets.

In the following tests, the criterion for a valid prediction is defined as at least one of the structural failures, such as spalling and cracks, being inside a bounding box or mask, although sometimes there are several bounding boxes or masks in an image when the threshold is low. Metrics including recall, precision and total accuracy are used for evaluating the performance of these models:

$$Recall = \frac{TP}{TP+FN} \tag{1}$$

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Accuracy = \frac{TP+TN}{TP+FP+FN+FN} \tag{3}$$

where *TP* and *TN* are true positive and negative, *FP* and *FN* are false positive and negative, respectively.

Methods	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Cascade Mask R-CNN	7.4	21.4	3.3	4.2	13.9	6.5
APANet Mask R-CNN	6.3	21.3	1.7	4.9	9.0	7.1
HRNet Mask R-CNN	5.9	19.9	2.1	5.0	6.9	6.8

Table 1. Comparison on Mask R-CNNs with Validation Data for Cracks.

Methods	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Cascade Mask R-CNN	20.3	39.0	19.2	1.2	16.7	23.2
APANet Mask R-CNN	13.9	33.0	10.7	0.7	12.2	16.1
HRNet Mask R-CNN	14.7	33.3	10.8	0.2	12.3	16.9

Table 2. Comparison on Mask R-CNNs with Validation Data for Spalling.

4.2 Tests on Phi-Net dataset (Gao and Mosalam, 2020)

In the dataset, the following classes are annotated for eight tasks: 1) scene levels; 2) damaged or undamaged states; 3) spalling or nonspalling; 4) material types; 5) collapse modes; 6) component types: including beams, columns, walls and others; 7) damage levels; 8) damage types. Totally, 36,413 images are collected in this dataset, but just for training and testing with classification. In addition, all images are low-resolution ones since the image size is uniformly resized as 224×224. From these images, we merged Task 3 which includes spalling and nonspalling cases and Task 8 which collects cracks and no cracks scenarios into a new testing dataset. The total number of the dataset is 5,853. The threshold for spalling and cracks being detected is set as 0.2 instead of 0.5 as most studies used. Figure 6 shows the examples of successful prediction by three Mask R-CNNs. In these overlaid images, the bounding boxes, and masks of spalling and cracks are in green, purple, and yellow colors, respectively. These colors have the same meaning in Figures 6, 8, 7, and 9.

Methods	Accuracy	Recall	Precision
Cascade Mask R-CNN	78.9%	70.7%	**88.7%**
APANet Mask R-CNN	**81.1%**	84.8%	83.8%
HRNet Mask R-CNN	58.6%	**95.5%**	57.7%

Table 3. Predictions of Mask R-CNNs on Phi-Net dataset.

From Table 3, it can be observed that the accuracy of Mask R-CNN + HRNet is quite low, but Cascade Mask R-CNN and APANet Mask R-CNN is higher over this low-resolution image dataset. Furthermore, both of them have a very high recall and precision.

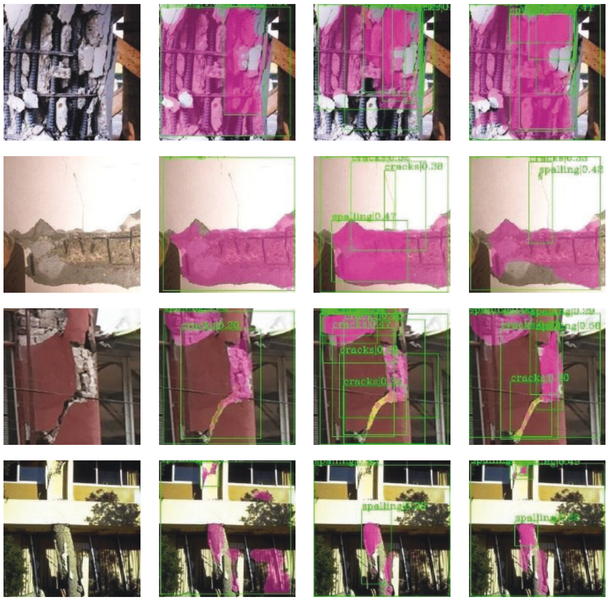

(a) (b) (c) (d)

Figure 6. Prediction of three Mask R-CNNs for Phi-Net dataset (Gao and Mosalam, 2020). (a), (b), (c), and (d) denote original image, overlaid image of Cascade, APANet, and HRNet Mask R-CNN, respectively.

4.3 Tests on 2017 Mexico City earthquake dataset (Purdue University, 2018)

In this dataset, there are 4,136 images with two image resolutions, 2740×3650 and 6000×4000. All of the images are taken by experts at Purdue University when they conducted the field investigation in Mexico City after a Richter magnitude 7.1 earthquake in 2017. Figure 7 shows some examples of correct prediction from our models.

The accuracy of APANet Mask R-CNN is higher than the others, and the accuracy of other two methods is close (see Table 4). The recall and precision of APANet Mask R-CNN and HRNet Mask R-CNN are above 73.0%, but the recall of Cascade Mask R-CNN is low.

Methods	Accuracy	Recall	Precision
Cascade Mask R-CNN	69.4%	45.5%	**90.9%**
APANet Mask R-CNN	**74.7%**	70.4%	85.7%
HRNet Mask R-CNN	69.1%	**73.5%**	73.4%

Table 4. Predictions of Mask R-CNNs on 2017 Mexico City earthquake dataset.

4.4 Tests on 2017 Pohang earthquake dataset (Sim et al., 2018)

In this dataset, a research group supported by the American Concrete Institute (ACI) collected images during their inspection after an earthquake with the Richter magnitude of 5.2 happened in Pohang of South Korea in 2017. The total number of images used for testing is 4,109, and their resolutions are 2600×3890 and 5180×3460. Some examples of good predictions are shown in Figure 8. The accuracy, recall and precision of three models are shown in Table 5.

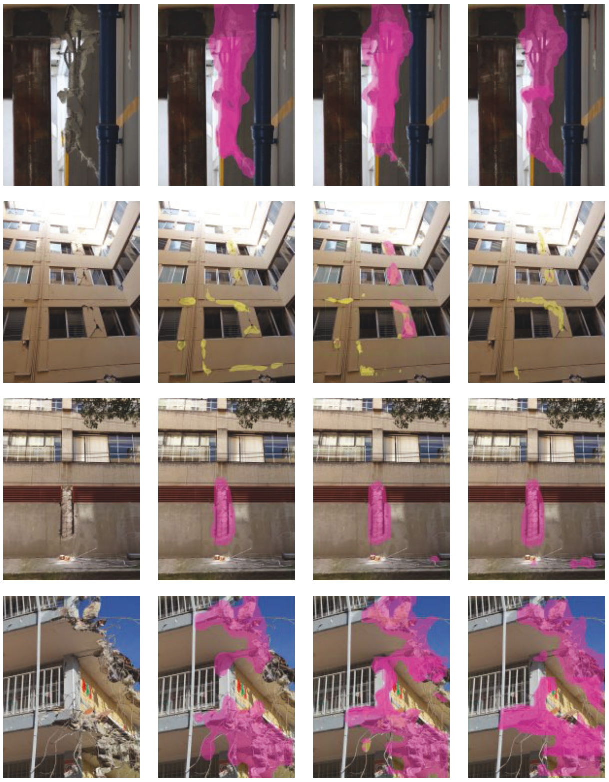

(a) (b) (c) (d)

Figure 7. Prediction of three Mask R-CNNs for 2017 Mexico City earthquake dataset (Purdue University, 2018). (a), (b), (c), and (d) denote original image, overlaid image of Cascade, APANet, and HRNet Mask R-CNN, respectively.

Methods	Accuracy	Recall	Precision
Cascade Mask R-CNN	66.0%	37.3%	**91.8%**
APANet Mask R-CNN	67.6%	57.6%	79.3%
HRNet Mask R-CNN	**68.1%**	**67.0%**	75.6%

Table 5. Predictions of Mask R-CNNs on 2017 Pohang earthquake dataset.

In Table 5, the accuracy of these three models is close to each other. Cascade Mask R-CNN have the highest precision. but its recall is quite low. The precision for APANet Mask R-CNN is also higher than HRNet Mask R-CNN while its recall is lower than the later in this dataset.

4.5 Failure cases

It should be noted that these models have been distracted by the crack-like or spalling-like objects during the tests. The major reason is that the training data are insufficient to cover all kinds of scenes when spalling and cracks appeared on the structures or its components are captured by the cameras. It is also a common problem for training and testing deep learning methods of instance segmentation. Some examples of wrong prediction for these three public datasets are shown in Figure 9. We associate the distractions as these: 1) wires or cables; 2) trees; 3) fences; 4) shadow; 5) edges of windows, buildings or other artifact objects.

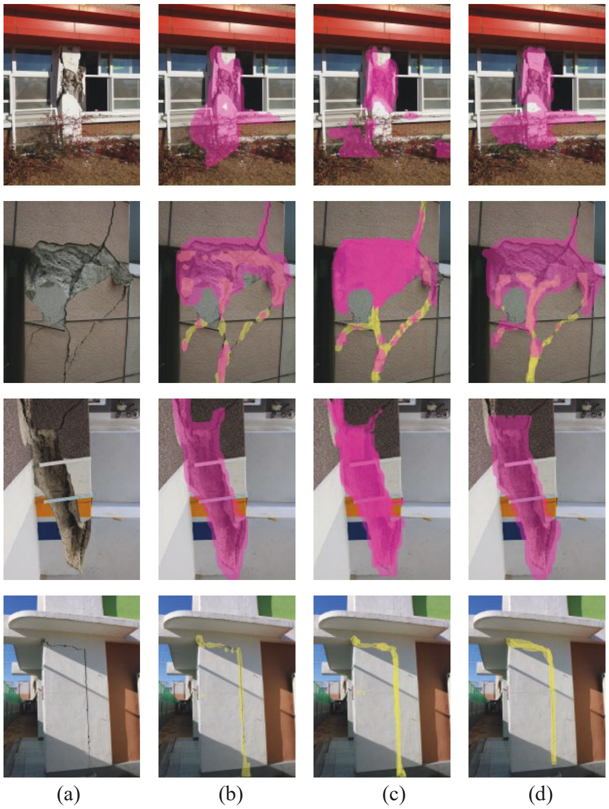


Figure 8. Prediction of three Mask R-CNNs for 2017 Pohang earthquake dataset (Sim et al., 2018). (a), (b), (c), and (d) denote original image, overlaid image of Cascade, APANet, and HRNet Mask R-CNN, respectively.

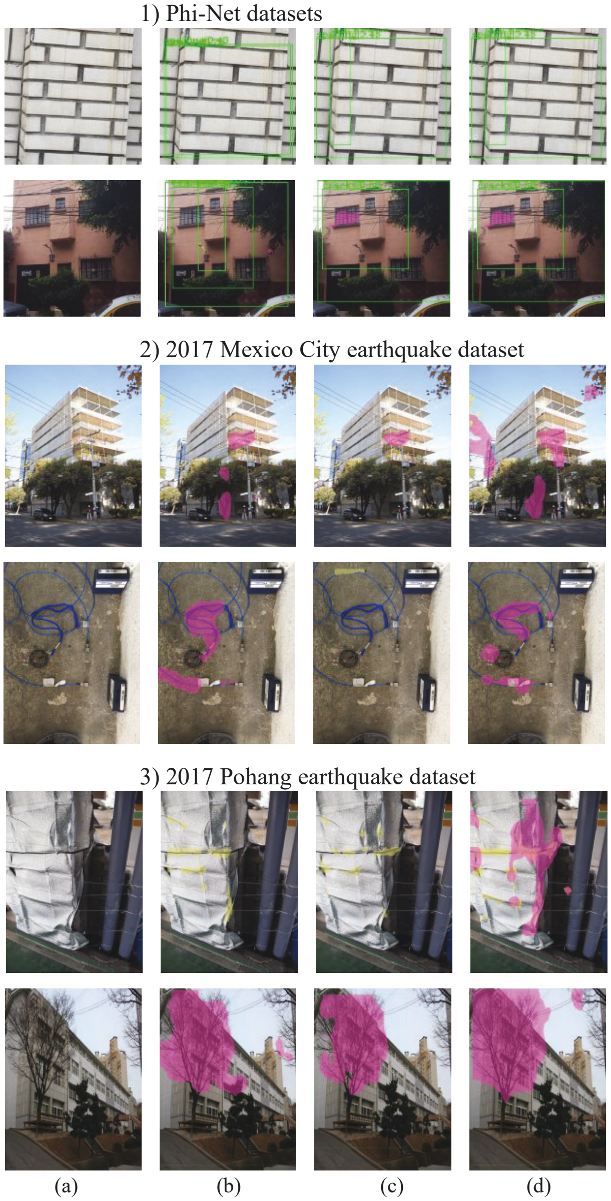
5. DISCUSSION

With an aim to find an end-to-end framework to detect cracks and spalling automatically and accurately, three Mask R-CNNs have been evaluated on three different public image datasets collected in different extreme events. In our analysis, the followings are observed:

1) Low resolution is commonly used with high-speed cameras whereas high resolution is standard for lower fps but high-quality data collection. Our testing results show that APANet Mask R-CNN can be a robust model to detect cracks and spalling with low- or high-definition images.

2) The scale of the scene in the image is also another important factor affecting the success rate to detect the structural damage, namely spalling and cracks. The models have higher accuracy when the cameras are closer to the damage, but they fail when the damage, especially cracks, are viewed from far and become invisible if the camera and the damage are so distant. In addition, the false predictions are very common when there are many crack-like or spalling-like objects at large scales. It may be solved through collecting more similar images and labelling them for training.

3) Compared to low resolution, implementation of these models takes longer time on high-resolution images. This is due to the increase in the number of pixel-wise processes with an increase in image size. We also found out that the masks of the models do not exactly fit the shapes and positions of these two types of structural damage in some cases. Furthermore, not every piece of cracks or spalling is marked separately. Exploring solutions for this problem is a future task.


1) Phi-Net datasets

2) 2017 Mexico City earthquake dataset

3) 2017 Pohang earthquake dataset

Figure 9. Some examples of wrong predictions of three Mask R-CNNs for three public datasets. (a), (b), (c), and (d) denote original image, overlaid image of Cascade, APANet, and HRNet Mask R-CNN, respectively.

4) This paper is a good showcase to apply the latest instance segmentation networks on detecting SDD for field investigations.

6. CONCLUSIONS

In this study, we tested three different Mask R-CNN architectures for detecting and segmenting cracks and spalling. Our goal is to show that these frameworks can be used as an end-to-end solution for the task independent of damage scales or image resolutions, which cause issues for instance segmentation of structural damage like cracks and spalling. Although the damage to the buildings and bridges in affected regions vary significantly in extreme events, the APANet Mask R-CNN was shown to achieve an accuracy above 67.6% for automatically detecting spalling and cracks on concrete and masonry structures. In the future, a more comprehensive dataset for better training will be made to quantify the damage and detect more types of structural failures while increasing the accuracy and precision of the damage position and boundary.

The link for the training and validation data of this study is here: https://github.com/OSUPCVLab/CrSpEE.

ACKNOWLEDGEMENTS

This study is based upon work supported by the U.S. National Science Foundation under grant no. 2036193.

REFERENCES

Ali, L., Khan, W., Chaiyasarn, K., 2019. Damage Detection and Localization in Masonry Structure Using Faster Region Con- volutional Networks. *International Journal OF GEOMATE*, 17(59), 98–105.

Atha, D. J., Jahanshahi, M. R., 2018. Evaluation of deep learning approaches based on convolutional neural networks for corrosion detection. *Structural Health Monitoring*, 17(5), 1110–1128.

Attard, L., Debono, C. J., Valentino, G., Di Castro, M., Masi, A., Scibile, L., 2019. Automatic crack detection using mask rcnn. *2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA)*, IEEE, 152–157.

Bai, Y., Yilmaz, A., Sezen, H., 2020a. End-to-end Deep Learning Methods for Automated Damage Detection in Extreme Events at Various Scales. *2020 25th International Conference on Pattern Recognition (ICPR)2021,* IEEE, 6640–6647.

Bai, Y., Zha, B., Sezen, H., Yilmaz, A., 2020b. Deep cascaded neural networks for automatic detection of structural damage and cracks from images. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, V-2-2020, 411– 417. https://www.isprs-ann-photogramm-remote-sens-spatialinf-sci.net/V-2-2020/411/2020/.

Brooks, J., 2019. COCO Annotator. https://github.com/jsbroks/coco-annotator/.

Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A. A., 2020. Albumentations: Fast and Flexible Image Augmentations. Information, 11(2). https: //www.mdpi.com/2078-2489/11/2/125.

Cai, Z., Vasconcelos, N., 2019. Cascade R-CNN: High Quality Object Detection and Instance Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1. http://dx.doi.org/10.1109/tpami.2019.2956516.

Cha, Y.-J., Choi, W., Suh, G., Mahmoudkhani, S., Büyüköztürk, O., 2018. Autonomous structural visual inspec-¨ tion using region-based deep learning for detecting multiple damage types. *Computer-Aided Civil and Infrastructure Engineering*, 33(9), 731–747.

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C. C., Lin, D., 2019. MMDetection: Open MMLab Detection Toolbox and Benchmark. *arXiv preprint arXiv:1906.07155.*

Gao, Y., Mosalam, K. M., 2018. Deep transfer learning for image-based structural damage recognition. *Computer-Aided Civil and Infrastructure Engineering*, 33(9), 748–768.

Gao, Y., Mosalam, K. M., 2020. New peer report 2019/07:” peer hub imagenet (Ø-net): A large-scale multi-attribute benchmark dataset of structural images”. https://peer.berkeley.edu/news/new-peer-report-201907-peer-hub-imagenet-Ø-net-large-scale-multi-attribute-benchmark-dataset.

Ghosh Mondal, T., Jahanshahi, M. R., Wu, R.-T., Wu, Z. Y., 2020. Deep learning-based multi-class damage detection for autonomous post-disaster reconnaissance. *Structural Control and Health Monitoring*, 27(4), e2507.

He, K., Gkioxari, G., Dollar, P., Girshick, R., 2017. Mask r-cnn.´ *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

Hoskere, V., Narazaki, Y., Hoang, T., Spencer Jr, B., 2018. Vision-based Structural Inspection using Multiscale Deep Convolutional Neural Networks. *arXiv preprint arXiv:1805.01055.*

Kalfarisi, R., Wu, Z. Y., Soh, K., 2020. Crack Detection and Segmentation Using Deep Learning with 3D Reality Mesh Model for Quantitative Assessment and Integrated Visualization. *Journal of Computing in Civil Engineering*, 34(3), 04020010.

Kim, B., Cho, S., 2019. Image-based concrete crack assessment using mask and region-based convolutional neural network. *Structural Control and Health Monitoring*, 26(8), e2381.

Kong, X., Li, J., 2018. Automated fatigue crack identification through motion tracking in a video stream. *Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2018*, 10598, International Society for Optics and Photonics, 105980V.

Liu, S., Qi, L., Qin, H., Shi, J., Jia, J., 2018. Path aggregation network for instance segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8759–8768.

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D., 2020. Image Segmentation Using Deep Learning: A Survey. *arXiv preprint arXiv:2001.05566.*

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 91–99.

Sim, C., Laughery, L., Chiou, T. C., Weng, P.-w., 2018. 2017 pohang earthquake - reinforced concrete building damage survey. https://datacenterhub.org/resources/14728.

Sun, K., Zhao, Y., Jiang, B., Cheng, T., Xiao, B., Liu, D., Mu, Y., Wang, X., Liu, W., Wang, J., 2019. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514.*

Purdue University, 2018. Buildings surveyed after 2017 mexico city earthquakes. https://datacenterhub.org/resources/14746.

Yang, L., Li, B., Li, W., Jiang, B., Xiao, J., 2018. Semantic metric 3d reconstruction for concrete inspection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 1543–1551.

Zha, B., Bai, Y., Sezen, H., Yilmaz, A., 2019. Deep Convolutional Neural Networks for Comprehensive Structural Health. *International journal of computer vision*, 115(3), 3367-3374.

Zhu, X., Cheng, D., Zhang, Z., Lin, S., Dai, J., 2019. An empirical study of spatial attention mechanisms in deep networks. *Proceedings of the IEEE International Conference on Computer Vision*, 6688–6697.