Fine-tuned Pre-trained Mask R-CNN Models for Surface Object Detection

Haruhiro Fujita  
Faculty of Management and Information Sciences  
Niigata University of International and Information Studies  
Niigata City, Japan  
fujita@nisis.ac.jp

Masatoshi Itagaki  
BSN iNet Co. Ltd.  
Niigata City, Japan  
masa-ita@bsn-net.co.jp

Kenta Ichikawa  
BSN iNet Co. Ltd.  
Niigata City, Japan  
ichikawa@bsn-net.co.jp

Yew Kwang Hooi  
Department of Computer and Information Sciences  
Universiti Teknologi PETRONAS  
Bandar Sri Iskandar, Malaysia  
yewkwanghooi@utp.edu.my

Kazutaka Kawano  
Department of Curatorial Research  
Tokyo National Museum  
Tokyo, Japan  
kawano4191kka489@gmail.com

Ryo Yamamoto  
Department of Curatorial Planning  
Tokyo National Museum  
Tokyo, Japan  
yamamoto-r62@nich.go.jp

Abstract—This study evaluates road surface object detection tasks using four Mask R-CNN models as a pre-study of surface deterioration detection of stone-made archaeological objects. The models were pre-trained and fine-tuned by COCO datasets and 15,188 segmented road surface annotation tags. The quality of the models were measured using Average Precisions and Average Recalls. Result indicates substantial number of counts of false negatives, i.e., “left detection” and “unclassified detections”. A modified confusion matrix model to avoid prioritizing IoU is tested and there are notable true positive increases in bounding box detection, but almost no changes in segmentation masks.

Keywords—road object detection, Mask R-CNN models, TensorFlow Object Detection API, mAP, AR, confusion matrices

I. INTRODUCTION

Computer vision is a broad science field handling various types of cognitive image processing including object detection [1]. Object detection is a task of finding different objects in an image and classifying them. Object detection is a critical task in autonomous driving where advanced deep learning models in conjunction with many AI sensors play major roles to control the system [2]. Various objects on the road surface including potholes and severe cracks require vehicle to identify and maneuver over them smoothly in real-time.

R-CNN (Region Convolutional Neural Network) is a very effective method for identifying objects. Results are shown as bounding box around detected objects with suitable class label. Ever since introduction of R-CNN for object detection in 2014, various improved algorithms have been introduced, such as Fast R-CNN, Faster R-CNN and Mask R-CNN.

R-CNN extracts region of proposals, compute the CNN features and classify the regions accordingly. Selective search uses windows of various dimensions to determine regions based on pixel features such as color, intensity and texture. The identified regions are warped into a standard square size and parsed to a Convolutional Neural Network for features extraction. Finally, Support Vector is used to classify the regions.

Although speed has improved through optimization of the learning process, speed remains a critical problem for real-time object detection for surface object detection from a moving sensor. Various existing CNN models can be studied for improving bounding box potentials for surface object detection.

This manuscript is organized as follows: literature review of surface object detection technologies and applications for road surface object detections; specific objectives of this study; methods of study; result of the testing the models; discussion of the profiles of each model and pinpointing critical areas for improvement.

II. LITERATURE REVIEW

A. Superpixel and Instance Segmentation

Superpixel is used as the base for instance segmentation, as a graphical preparatory clustering method. It clusters pixels of vicinity in terms of geometric and color spaces prior to object segmentation using SLIC (Simple Linear Iterative Clustering) algorithm [3, 4]. This study uses superpixel as the base for instance segmentation. It is defined and enhanced by the Common Object in Context (COCO). COCO is a large image-based datasets with object segmentation, recognition in context, differentiation of objects (cars, people, houses) and stuffs (roads, sky, wall), 200k labeled images, 1.5 million object instances, 80 object categories, 91 stuff categories, and 5 captions in each image [5].

Instance segmentation detects an object under overlap of other objects precisely. However, it does not label all of the pixels of the image, as it segments only the region of interests [6]. The state-of-the-art instance segmentation approach is the detection-based method that predicts the mask for each region after acquiring the instance region [7].

B. Mask Region Convolutional Neural Network (Mask R-CNN)

Fast R-CNN, introduced in 2015 to improve the speed of R-CNN, is a complex pipeline. It requires huge number of forward passes and separate training of three models for image
features generation, classification and regression respectively. Faster R-CNN improves Fast R-CNN for less complex training through ROI Pooling, a technique that limits the number of unnecessary forward passes through shared computation result. However, Faster R-CNN is limited to detect object in bounding boxes only.

Enters Mask Region Convolutional Neural Network (Mask R-CNN), as proposed by [8]. Mask R-CNN is an extended model of the Faster R-CNN [9] which provides better detection of objects in segmentations. Mask R-CNN adds learning segmentation masks (called Mask Branch) sub-module onto Faster R-CNN. The sub-module predicts segmentation masks on each Region of Interest (RoI) [6] in parallel with each other using convolution arrays of data for classification and bounding box regression. The mask module is a small fully convolutional network (FCN) applied to each RoI, predicting a segmentation mask for each pixel.

C. Four Mask R-CNN models in TensorFlow Object Detection API

TensorFlow Object Detection API is an open archive of deep learning models for object detection. It contains pre-trained models with multiple datasets including COCO. The pre-trained models generate bounding boxes around objects. From the various models, four Mask R-CNN models which generate masks (instance segments) of identified objects are selected for this study [10].

D. Road damage detection system and models

An application of object detection is road damage detection. The first reported road surface object detection models using Mask R-CNN are cited in [11], as shown Fig. 1.

![Fig. 1. Road surface object detection using one of Mask R-CNN models](image)

Another related work is that of Maeda et al.[12] who had trained Single Shot Detection (SSD) models on their dataset captured with Smartphones.

III. OBJECTIVES

The objectives of this study are to:-

- evaluate several pre-trained Mask R-CNN models in TensorFlow Object Detection API
- fine-tune the models by annotation data of road objects
- study the applicability of the models to the road object detection

IV. METHODOLOGY

The methodologies are the same as previously reported in [11]. To optimize performance of road surface detection, an enhanced method was applied to train pre-learned four Mask R-CNN models in TensorFlow Object Detection API. The models are then evaluated using sorted annotation test data.

A. Data and Model Preparation

Refer to TABLE I. The annotation data contains 12 classes of road damages and cognitive objects in the road images to differentiate those damages from others. Microsoft Visual Object Tagging Tool (VoTT) was used for annotating 1,418 road color images with a size of 3840 x 2160. All visible objects of the predefined classes in two third bottom part of each image were segmented and tagged.

| classes          | tag      | datasets/segmentations | training | validation | testing | total |
|------------------|----------|------------------------|----------|------------|---------|-------|
| Linear Cracks    | Crack1   | 2,315                  | 468      | 455        | 3,238   |
| Grid Cracks      | Crack2   | 524                    | 103      | 101        | 728     |
| Pavements joints | Joint    | 943                    | 158      | 159        | 1,358   |
| Patchings        | Patching | 377                    | 102      | 77         | 556     |
| Fillings         | Filling  | 1,240                  | 207      | 187        | 1,634   |
| Pot-holes        | Pothole  | 145                    | 39       | 14         | 198     |
| Marks/holes      | Markhole | 296                    | 61       | 52         | 409     |
| Stains           | Stain    | 96                     | 22       | 12         | 132     |
| Shadow           | Shadow   | 1,084                  | 252      | 212        | 1,548   |
| Pavement Markings| Marking  | 1,239                  | 241      | 297        | 1,777   |
| Scratches on Markings | Scratch | 2,433                  | 414      | 576        | 3,503   |
| Grid Crack in Patchings | Patching | 73                    | 19       | 17         | 109     |
| Total            |          | 10,763                 | 2,209    | 2,219      | 15,188  |

The most numerous segments were “Scratches on Markings” class at total number of 3,503 segments. This is followed by “Linear Cracks” at 3,238 segments. The minimum segments were “Grid Cracks in Patchings” at 109 segments followed by “Stains” and “Pot-holes”. The segments of each class were divided into datasets for training, training validation (validation) and testing at ratio of 0.7:0.15:0.15.

B. Computation Resources

The tests were conducted on a machine with Intel Core i7-8700K (6 cores/12 threads/3.70GHz) CPU, 32GB CPU memory, Nvidia GeForce GTX 1080 GPU with 2560 CUDA core, GPU clocks of 1607/1733 MHz, 8GB GDDR5X GPU memory and 320GB/s memory bandwidth. The language for programming is Python 3.6.9 using libraries from Object Detection API version 1 on top of TensorFlow 1.15.0.

C. Methods

Main steps of the methods are performing annotation, mask data making; model training and fine tuning.

1) Annotation and Mask Data Making:

Refer to Fig. 2. The files output by the Visual Object Tagging Tool (VoTT) were converted to intermediate annotation files. Next, the files were converted into TensorFlow Record files for training, validation and testing. The conversions were done using three separate conversion script executive programs in Python. The last conversion script (vott2tfrecords.py) modified the original image size of
3840x2160 (width x height) to match with the training image size of 960x540 used in machine learning.

2) Model Training and Fine Tuning:
Refer to Fig. 3. Pre-trained model checkpoint file, TensorFlow Record files, ground truth annotation data/labels (“tf_label map.pbtxt”) and pipeline.config are used for training of object detection. The pretrained model checkpoint is based on MS COCO dataset. The TensorFlow Record files contain both datasets for training and validation. The training generated Fine Tuned Checkpoint file from which an object detection model was acquired. Finally, the fine-tuned detection model was evaluated using validation and testing datasets. The inference result produces precision metric, recall metric and confusion matrix. The training was done for each four Mask R-CNN models.

D. Model Evaluation Metrics
Various metrics used for model evaluation of this work are mean average precisions (mAP) and average recalls (AR) at various levels of Intersection on Unit(IoU). The Mask R-CNN has a region proposal network layer that makes multiple inferences simultaneously on the class classification, the segmentation and the mask areas resulting six loss metrics.

Besides above model-wise metrics, the average precisions and the average recalls at IoU=0.5 are used for all twelve road object classes.

TABLE II: MODEL METRICS OF MASK R-CNN INCEPTION V2 COCO ON BOUNDING BOX

| Tag          | Precision | Recall |
|--------------|-----------|--------|
| Crack1       | 0.4058    | 0.2549 |
| Crack2       | 0.3623    | 0.3881 |
| Joint        | 0.5446    | 0.4286 |
| Filling      | 0.4715    | 0.3200 |
| Pothole      | 0.5174    | 0.2827 |
| Manhole      | 0.0200    | 0.7000 |
| Stain        | 0.3975    | 0.3019 |
| Marking      | 0.0224    | 0.0312 |
| Scratch      | 0.0504    | 0.0223 |
| Patching1    | 0.4000    | 0.1176 |

TABLE III: MODEL METRICS OF MASK R-CNN INCEPTION V2 COCO ON SEGMENTATION MASK

| Tag          | Precision | Recall |
|--------------|-----------|--------|
| Crack1       | 0.3409    | 0.1200 |
| Crack2       | 0.4502    | 0.5496 |
| Joint        | 0.1320    | 0.1215 |
| Filling      | 0.1142    | 0.0245 |
| Pothole      | 0.2574    | 0.2877 |
| Manhole      | 0.6706    | 0.7920 |
| Stain        | 0.0000    | 0.0000 |
| Shadow       | 0.1387    | 0.1415 |
| Marking      | 0.2107    | 0.2196 |
| Scratch      | 0.3505    | 0.3945 |
| Patching1    | 0.6090    | 0.1795 |

![Fig. 3. Training and evaluation processes of pre-trained Mask R-CNN models.](image)

![Fig. 2. Flow for annotation and data analysis.](image)

The detection precisions at IoU=.50 found in our target damage classes of linear cracks (Crack1), grid cracks (Crack2), pot-holes, scratches on markings, grid cracks in patchings are 0.4085, 0.4958, 0.5714, 0.5934 and 0.4000 respectively.

V. RESULTS

A. Evaluation Metrics of Mask R-CNN variants

1) Inceptions v2 COCO:
TABLE II and TABLE III shows various metrics of the model on bounding boxes and segmentation masks respectively. The values of mAP(IoU=50.05.95), mAP(IoU=.50) and mAP(IoU=.75) are 0.2432, 0.4382 and 0.2482 on bounding boxes, and 0.1600, 0.3257 and 0.1279 respectively, with notable reduction in those metrics on segmentation masks. The Precision mAP(small) for small objects = 0.0365 and 0.0133 on bounding boxes and segmentation masks respectively are notably small, compared with those Precision mAP(large) of large objects and medium objects. The Average Recall for small, medium and large object are 0.1166, 0.3132 and 0.4717 on boundary boxes and 0.1021, 0.2528 and 0.2732 on segmentation masks respectively.

The detection precisions at IoU=.50 found in our target damage classes of linear cracks (Crack1), grid cracks (Crack2), pot-holes, scratches on markings, grid cracks in patchings are 0.4085, 0.4958, 0.5714, 0.5934 and 0.4000 respectively.

![Fig. 4 shows the model's confusion matrix of the model on bounding boxes (left) and segmentation masks (right) respectively.](image)
For example, in the left matrix of the bounding boxes, in the first horizontal row from the left, 116 is the count of true class of “Crack1” and predicted class of “Crack1” or linear cracks. It indicates the number of boundary boxes of true class linear cracks predicted correctly. Other cells in the same row indicates “Crack1” wrongly predicted as other classes.

The last count 329 is the number of boundary boxes where ground truths of linear crack existed but is not detected in any classes. This means the model could not predict ground truths as any classes”, shorty “left detection” or False Negative (F/N). The bottom row are the “False Negative” counts, i.e. the boundary boxes that were detected but not classified at any classes for lower IoU values. Vice versa in the confusion matrix of segmentation masks.

Substantial “left detection” counts are found in linear cracks, grid cracks, joints, patchings, fillings, potholes, stains, shadows, patchings with grid cracks, where the counts exceed 50% of the correct prediction counts, both in bounding boxes and segmentation masks. Substantial “unclassified detections” (where counts exceed 50% of True Positive) are found in linear cracks, grid cracks, joints, fillings, grid cracks, shadows and fillings with grid cracks, both in bounding boxes and segmentation masks.

2) Inception ResNet v2 Atrous COCO:

TABLE IV and TABLE V show mAP(IoU=0.50:.05:.95), mAP(IoU=0.50) and mAP(IoU=0.75) of the model, which are 0.2756, 0.4678 and 0.2815 on bounding boxes, and 0.2096, 0.4038 and 0.1982 on segmentation masks respectively. The detection precisions at 0.3571, 0.5321, 0.5455, 0.5937 and 0.5780, 0.4328, 0.6391 and 0.6268 on boundary boxes and grid cracks, potholes, scratches and markings are 0.3791, 0.5780, 0.4328, 0.6391 and 0.6268 on boundary boxes and grid cracks. Precision mAPs (medium) = 0.1852 and 0.1552 in bounding boxes and segmentation masks respectively. The detection precisions at 0.3571, 0.5321, 0.5455, 0.5937 and 0.5780, 0.4328, 0.6391 and 0.6268 on boundary boxes and grid cracks, joint, fillings, potholes, scratches and markings are 0.1552, 0.3562 and 0.3325 on small, medium and large object are 0.1852, 0.3562 and 0.3325 on small, medium and large object. The confusion matrix of segmentation masks.

In Fig. 5 confusion matrices on bounding boxes (left) and segmentation masks (right). They indicate substantial “left detection” (which exceeds 50% of True Positive) in all classes, except for grid cracks (bounding boxes only), manholes, markings (bounding boxes only) and scratches (bounding boxes only). The substantial “unclassified detections” counts (which exceeds 50% of True Positive) are found in linear cracks, grid cracks, joints, fillings, stains, shadows and fillings with grid cracks.

3) Inception ResNet50 Atrous COCO:

See Fig. 5 confusion matrices on bounding boxes (left) and segmentation masks (right). They indicate substantial “left detection” (which exceeds 50% of True Positive) in all classes, except for grid cracks (bounding boxes only), manholes, markings (bounding boxes only) and scratches (bounding boxes only). The substantial “unclassified detections” counts (which exceeds 50% of True Positive) are found in linear cracks, grid cracks, joints, fillings, stains, shadows and fillings with grid cracks.

See Fig. 5 confusion matrices on bounding boxes (left) and segmentation masks (right).

TABLE IV. MODEL METRICS OF MASK R-CNN INCEPTION RESNET V2 ATROUS COCO ON BOUNDING BOXES

| IoU | Precision | Recall |
|-----|-----------|--------|
| 0.50 | 0.3571    | 0.2096 |
| 0.55 | 0.5321    | 0.4038 |
| 0.60 | 0.5455    | 0.1982 |
| 0.65 | 0.5937    | 0.3562 |
| 0.70 | 0.5780    | 0.6391 |
| 0.75 | 0.4328    | 0.6268 |

TABLE V. MODEL METRICS OF MASK R-CNN INCEPTION RESNET V2 ATROUS COCO ON SEGMENTATION MASKS

| IoU | Precision | Recall |
|-----|-----------|--------|
| 0.50 | 0.3571    | 0.2096 |
| 0.55 | 0.5321    | 0.4038 |
| 0.60 | 0.5455    | 0.1982 |
| 0.65 | 0.5937    | 0.3562 |
| 0.70 | 0.5780    | 0.6391 |
| 0.75 | 0.4328    | 0.6268 |

See Fig. 5 confusion matrices on bounding boxes (left) and segmentation masks (right).
mAP(IoU=.75) are 0.2810, 0.4789 and 0.2965 on bounding boxes, and 0.2104, 0.4140 and 0.1934 on segmentation masks respectively. The mAPs for small objects of 0.0800 and 0.0452 on bounding boxes and segmentation masks are more than twice and three times as much as the same metric of Mask R-CNN Inception v2 COCO. Notable reductions in mAPs and ARs, however their magnitudes are much smaller in ARs. A notable smaller detection precisions at IoU=.50 on segmentation masks are found in those classes of linear cracks, joints, fillings, potholes, stains, shadows, scratches (only on segmentation masks) and patchings with grid cracks.

4) Inception ResNet101 Atrous COCO:

TABLE VIII and TABLE IX show Inception ResNet101 Atrous COCO’s mAP(IoU=.50:05:.95), mAP(IoU=.50) and mAP(IoU=.75), which are 0.2779, 0.4787 and 0.2853 on bounding boxes, and 0.2118, 0.4036 and 0.2045 on segmentation masks respectively, the latter show the highest precision among those four models. Average Recalls are 0.4975, 0.3759 and 0.1665 on bounding boxes and 0.3404, 0.3382 and 0.1475 on segmentation masks for large, medium and small objects respectively.

Confusion matrices in Fig. 6, indicate substantial “left detection” as noted in the previous model. This is observed on bounding boxes in linear cracks, joints, fillings, stains, shadows, and patchings with grid cracks (left detection counts exceed the correctly predicted). The substantial “unclassified detections” counts (which exceed 50% of True Positive) are found in, linear cracks, grid cracks, joints, fillings, potholes, stains, shadows, scratches (only on segmentation masks) and patchings with grid cracks.

| Precision | Recall |
|-----------|--------|
| **cracks** | precision | recall |
| Crack1 | 0.022 | 0.107 |
| Crack2 | 0.052 | 0.263 |
| Crack3 | 0.022 | 0.107 |
| Crack4 | 0.022 | 0.107 |
| Crack5 | 0.022 | 0.107 |
| Crack6 | 0.022 | 0.107 |
| Crack7 | 0.022 | 0.107 |
| Crack8 | 0.022 | 0.107 |
| Crack9 | 0.022 | 0.107 |
| Crack10 | 0.022 | 0.107 |
| Crack11 | 0.022 | 0.107 |
| Crack12 | 0.022 | 0.107 |
| Crack13 | 0.022 | 0.107 |
| Crack14 | 0.022 | 0.107 |
| Crack15 | 0.022 | 0.107 |
| Crack16 | 0.022 | 0.107 |
| Crack17 | 0.022 | 0.107 |
| Crack18 | 0.022 | 0.107 |
| Crack19 | 0.022 | 0.107 |
| Crack20 | 0.022 | 0.107 |

Fig. 7 confusion matrices of InceptionResNet101 Atrous COCO model indicate substantial “left detection” as noted in the previous model. This is observed on bounding boxes in linear cracks, joints, fillings, stains, shadows, and patchings with grid cracks (left detection counts exceed the correctly predicted), while patching is more than 50% of True Positive. Substantial “false unclassified detections” counts (which exceed 50% of True Positive) are also found on bounding boxes and segmentation masks in linear cracks, grid cracks, joints, fillings, potholes, stains, shadows, scratches (only on segmentation masks) and patchings with grid cracks.
there were substantial number of counts for left detection, i.e. “the model could not fit ground truths with detected objects”. This can be seen in the last horizontal row of each confusion matrix, there were substantial number of “unclassified detections” in many classes as well.

5) Result Summary of TensorFlow Object Detection API Models

Comparing the results obtained in all four pre-trained Mask R-CNN models, mask R-CNN resnet101 atrous coco (model four) has shown the best precisions and superior mean Average Precisions (small), Average Recalls@100, Average Recalls@10 and Average Recalls@100(small). mAPs generally reached the quick peak, but ARs gradually increased.

From the confusion matrices of all four models, there are substantial number of counts for left detection, i.e. “the model could not fit ground truths with detected objects”. This can be seen in the last horizontal row of each confusion matrix, there were substantial number of “unclassified detections” in many classes as well.

B. APs in training progresses of four models

The progress of mAPs of those four models up to 100k steps are shown in Figure 8. The model four showed the advantageous performances in mAP(IoU=.50:.05:.95), and in other five mAP metrics, the faster initial precision increases and the sustaining higher precisions until 100k steps. The model two showed better precision in mAP(IoU=.75) from 50k to 80k steps, than the model four, however, the precision of the model four reached very close to that of model two at 100k steps. The significant higher mAP for small objects (the bottom right graph) of model four is notable. Although the model two reached the higher precision at 100k steps in several metrics, the model four showed better overall performances than other models. On the other hand, the model one showed the least performances in all six AP metrics.

C. ARs in training progresses of four models

All six Average Recall metrics (the six graphs) showed the significant higher recall rate progresses of the model four at most learning steps (Figure 9). The faster performance increases and the stable plateaus in all six AR metrics of the model four are notable. The overwhelming higher recalls of the model four to other models in six AR metrics were prominent compared with mAPs, where some of those mAP metrics were very similar to the model two, as previously noted. A special attention was drawn on much higher AR@100 for the small objects. Although the AR performances of the model one is similar to those in Aps, there is a notable less performance of the model two in AR@100 for the small objects.
D. Losses in training progresses of four models

Six loss metrics were generated as shown in Figure 10. There is a generic tendency in all six losses metrics that all four models showed the initial quick drops at 10k-20k followed by the gradual linear increases as steps increase. The gradients of the linear increases vary among those four models. Notable higher total loss of the model four is found.

E. Comparison of mAPs/ARs per class

The AP/AR@0.5IoU of those twelve classes are shown in Figure 11 and 12. The large fluctuations in AP@0.5IoU of pot-holes, stains, and grid cracks in patches are found. After the initial changes, APs@0.5IoU of patchings, fillings, grid cracks, manholes, shadows, markings and scratches are stable. A slight moderate decrease of APs@0.5IoU in linear cracks and joints, and the moderate increase in shadow are found. Compared with those APs@0.5IoU, those metrics of ARs@0.5IoU in all twelve classes are very stable, not much fluctuations as seen in APs@0.5IoU. Except for grid cracks in patchings, all ARs@0.5IoU in other eleven classes are moderately increased.
Fig. 11. AP(IoU=.50) and AR(IoU=.50) of twelve road surface object classes (Part 1)
| ID | class   | precision@0.5IOU | recall@0.5IOU |
|----|---------|------------------|---------------|
| 7  | Manhole |                  |               |
| 8  | Stain   |                  |               |
| 9  | Shadow  |                  |               |
| 10 | Marking |                  |               |
| 11 | Scratch |                  |               |
| 12 | Patching2 |               |               |

Fig. 12. AP(IoU=0.5) and AR(IoU=0.5) of twelve road surface object classes (Part 2)
VI. DISCUSSION

A. Comparison of four Mask R-CNN models

For easier model identification, they are named as follows:

- Model #1: Mask R-CNN Inception v2 coco;
- Model #2: Mask R-CNN Inception ResNet v2 Atrous coco;
- Model #3: Mask R-CNN Inception ResNet50 Atrous coco;
- Model #4: Mask R-CNN Inception ResNet101 Atrous coco.

It is of interest to describe the performance of the four models before and after the road annotation training. Refer to TABLE X. After 200k labeled image training of COCO datasets, the highest mAPs (IoU=0.50:0.05:0.95) is found in model #2, with value of 0.36. This is followed by model #4 at value 0.33. Model #3 achieves value of 0.29 and the least performing model is model #1 with value 0.25. Substantial differences are observed in terms of detection performance among the four Mask R-CNN models.

| Model                          | Pre-trained model on COCO Dataset | Fine-tuned model with Road Dataset |
|-------------------------------|-----------------------------------|-----------------------------------|
| mask_rccn_inception_resnet_v2 | 0.36                              | 0.29                              |
| mask_rccn_inception_resnet_v2 | 0.25                              | 0.25                              |
| mask_rccn_inception_resnet50  | 0.29                              | 0.26                              |
| mask_rccn_inception_resnet101 | 0.33                              | 0.28                              |

After relatively small dataset training of those road annotation under 12 classes, the differences in mAP performances are much smaller than those without road object training. The highest was recorded for model #2, but the difference with the second highest model #4 is negligible, where the overall performance is much higher in model #4. The performance reduction in the model #3 and model #1, were also not as much as found in performances without road annotation training.

The detection precision metrics of mAP (IoU=0.50:0.05:0.95) were smaller in those models with road annotation data training, compared with those without training of road annotation data, although it is not adequate for making a direct comparison. However, those objects of the twelve classes on the road surfaces are not as much cognitive as those in the COCO datasets, in human eyes. Therefore, it is reasonable to conclude that the performance of models by road annotation training is not so bad, if one considers the severity of mAP (IoU=0.50:0.05:0.95), which is the primary challenge metric in COCO [13].

The overall model architectures are almost the same in all four models, with a difference in Feature Extractor. Model #1 can be classified as an Inception model, and other three models are combination of Inception and Resnet, where Residual Connection was added on top of Inception architecture [14].

It is notable that strides of Feature Extractor and Anchor Generator of model #1, are both 16, and those of other models are 8. It is also noted that the initial crop size of the Region Proposal Network in model #2 is 17, and those of other models are 14. There is a tendency that smaller parameter size of above modules brings capability to detect smaller objects. Those parameters are the same in model #3 and model #4. However, higher layer numbers in model #4 is advantageous than smaller layer numbers in model #3.

Therefore, it can be summarized that small number of strides of Feature Extractor and Anchor Generator, as well as smaller numbers of initial size of the Region Proposal Network, with more layers, made model #4 the best performance with special advantage on small object detection.

In the confusion matrices of all four models, there are substantial number of counts of no class prediction, namely “leaving judgement” or false negatives. In the last horizontal row there are substantial number of counts which have no class definition, although the horizontal row should be the true class, or ground truth. This enigma might be associated with threshold settings which makes direct influences on cut off of boundary boxes counts during the calculation.

B. IoU prioritized calculation problems of confusion matrix

Confusion matrix is a cross table of class detection counts in column, and class ground truth counts in rows. The diagonal cells of each corresponding class are true positives, i.e., correctly detected and classified instances. The bottom end of each column is counts of false positives, and right end of each row is counts of false negatives.

However, there is a fundamental problem of confusion matrix for regional object detection tasks, of which models are supposed to detect both a correct region and a correct object class simultaneously. As regional detection always sets an IoU (Interception over Unit) threshold of overlap extent of detected and ground truth areas, both in the bounding boxes and the segmentation masks. Therefore, the “correct detection” varies upon threshold variable. This study adopts IoU=0.5 as both thresholds for AP and AR metrics according to the COCO detection standard.

Refer Fig. 12 for IoU examples. The red color is a ground truth area, and the green color is a detected area at each IoU level. The COCO standard’s IoU=0.5 requires a substantial overlap extent in ground truth and detected areas.

However, some of regional object detection tasks do not require such a high standard, where an overlap extent of IoU=0.3 or IoU=0.2 might be quite enough, hence, the “regional correct detection” does not have a uniform standard.

The most commonly used regional detection algorithm created by Santiago Valdarrama, prioritizing the IoU conducts the following steps [15, 16].
calculates IoU with ground truths and detected areas, and makes a list of over threshold detection candidates

lists in IoU decent order

if there are several detection candidates over a single ground truth, discard all candidates by leaving one candidate with a maximum IoU

lists the remaining in decent order

if a single detection candidate matches multiple ground truths, discard all candidates by leaving one candidate with a maximum IoU

check the ground truth one by one

if a ground truth matches any detected candidates, count up it in a confusion matrix cell of a corresponding ground truth and detection class (true positive and false positive)

if a ground truth does not match any detected candidates, count up it in the right end cell of corresponding ground truth class (false negative)

check the detection candidates one by one

if one candidate does not match any ground truth, count up it in the bottom row of a corresponding detection class (false positive)

Our modified confusion matrix algorithm created by Masatoshi Itagaki is as followed.

Sort detected candidates in descend confidence score order

Pick one ground truth, if it has previously matched detected candidates pick next ground truth

Pick one detected candidate and calculate IoU

If a ground truth matches with detected candidates of the same class, select one with highest IoU as a match

If a ground truth has no detected candidates with the same class, select one of other classes with highest IoU as a "False Positive" match

If a detected candidate matches with more than a single ground truth instance, select a ground truth of higher IoU prioritizing the class

Repeat 2 to 7 until no more changes take place

Matches are counted up in corresponding class ID cells of the confusion matrix (as true positives and false positives)

If a ground truth has no matching detected candidates, count up right end column cells of corresponding ground truth class as false negatives

If a detected candidate has no ground truths matched, count up at the bottom row cells of corresponding class as false positives

The comparison of conventional and modified calculation methods is made at thresholds of IoU=0.5 and Confidence Score=0.5, for bounding boxes, and segmentation masks. The results are shown as in Figure 14 and Figure 15.

The true positive counts increased in most classes shown as in the diagonal cells of the modification (right graph) is notable. The slight decreases are found in false positives (the remaining counts from the true positive and the right end column counts). The false negative counts at the right end column, which are those left over ground truth counts from detection at those thresholds, shows few count changes, except for the scratch class, of ten counts reduction. The modified calculation method seems to reducing unnecessary rejection of true positives and false positives.

Refer Table XI. The notable increases in both precision and recall except for patching, pothole, manholes and stains classes.

The same comparisons of the conventional and the modified methods for segmentation masks in confusion matrices and class metrics are also made as in Figure 15 and Table XII. The results show almost no changes between the calculation methods, both in the confusion matrices and the class metrics. Very slight increases in metrics are found in grid crack, filling, marking and scratch classes but very slight decreases in linear crack and shadow classes.

In the modified calculation algorithm, if a ground truth does not match detection candidates with less class threshold IoU, but above threshold IoU in another class, it will be counted as a false negative. The less performance On segmentation masks by the modified calculation might be associated with smaller sizes of segments than bounding boxes, which leads less cases of above false negative counts.

![Image](image-url)
TABLE XII. CLASS METRICS BY CONVENTIONAL(LEFT), MODIFIED(RIGHT) CONFUSION MATRICES OF SEGMENTATION MASKS

| category     | precision | recall | category | precision | recall |
|--------------|-----------|--------|----------|-----------|--------|
| Crack1       | 0.3300    | 0.2800 | Crack1   | 0.3300    | 0.2800 |
| Crack2       | 0.5038    | 0.6014 | Crack2   | 0.5113    | 0.6733 |
| Joint        | 0.2732    | 0.2146 | Joint    | 0.2775    | 0.2182 |
| Patching     | 0.6308    | 0.5225 | Patching | 0.6106    | 0.5125 |
| Filling      | 0.3182    | 0.2620 | Filling  | 0.3247    | 0.2674 |
| Pushhole     | 0.5000    | 0.5000 | Pushhole | 0.5000    | 0.5000 |
| Manhole      | 0.1111    | 0.7885 | Manhole  | 0.1111    | 0.7885 |
| Stain        | 0.1538    | 0.1507 | Stain    | 0.1538    | 0.1647 |
| Shadow       | 0.3143    | 0.2594 | Shadow   | 0.3056    | 0.2694 |
| Marking      | 0.6400    | 0.5307 | Marking  | 0.6550    | 0.5522 |
| Scratch      | 0.5625    | 0.6146 | Scratch  | 0.5601    | 0.6233 |
| Patching2    | 0.5000    | 0.2353 | Patching2| 0.5000    | 0.2353 |

VII. CONCLUSION

This paper reports the final evaluation results obtained by test datasets and modified confusion matrix methodologies. There are substantial number of counts of left detection, as well as unclassified detections.

A modified confusion matrix model to avoid prioritizing IoU is tested. There are notable true positive increases in bounding box detection, but almost no changes found in segmentation masks.

The less performance on segmentation masks by the modified calculation might be associated with smaller sizes of segments than bounding boxes. This necessitates a further research.

ACKNOWLEDGMENT

Authors owe to Niigata University of International and Information Studies for funding support on the “Mobile Sensing Development by Deep Learning” project in 2020.

Special thanks are due to Fukuda Road Co. Ltd for its provision of road images for the annotation works as well as for the provision of information on Multi Fine Eye.

Finally Authors owe much to Tsuyoshi Ikenori for his works on the annotation data.

REFERENCES

[1] Szeliski, R.: ‘Computer Vision: Algorithms and Applications’ (Springer, 2010, 2010)
[2] Kaehler, A., and Bradski: ‘OpenCV 3’ (O’Reilly, 2016, 2016)
[3] Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Säärström, S.: ‘SLIC Superpixels Compared to State-of-the-art Superpixel Methods’, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34, (11), pp. 2274 – 2282
[4] https://techblog.shn-techorus.com/archives/7793
[5] https://cocodataset.org/
[6] Elenenbroich, C., Kutz, O., and Sattler, U.: ‘A Case for Abductive Reasoning over Ontologies’. Proc. Conference: Proceedings of the OWLED*06 Workshop on OWL: Experiences and Directions, Athens, Georgia, USA2006 pp. Pages
[7] https://engineer.dena.com/posts/2019.08/cv-papers-19-segmentation/
[8] He, K., Girshick, R., Dollar, P., and Girshick, R.: ‘Mask R-CNN’, in Editor (Ed.) (Eds.): ‘Book Mask R-CNN’ (Cornell University, 2017, edn.), pp.
[9] Ren, S., He, K., Girshick, R., and Sun, J.: ‘Faster R-CNN: Towards real-time object detection with region proposal net-works’, in Editor (Ed.) (Eds.): ‘Book Faster R-CNN: Towards real-time object detection with region proposal net-works’ (Cornell University, 2015, edn.), pp.
[10] https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/tf1_detection_zoo.md
[11] Haruhiro Fujita, Masatoshi Itagaki, Kenta Ichikawa, Yew Kuwan Hooi, Kazuyoshi Kawahara and Aliza Sarlan, Fine-tuned Surface Object Detection Applying Pre-trained Mask R-CNN Models, International Conference of Computational Intelligence 2020, IEEE Conference Proceeding, 978-1-5386-5541-2/18
[12] Maeda, H., Sekimoto, Y., Seto, T., Kashiyama, T., & Omata, H. Road Damage Detection and Classification Using Deep Neural Networks with Smartphone Images. Computer-Aided Civil and Infrastructure Engineering.
[13] https://blog.zenggyu.com/en/post/2018-12-16-an-introduction-to-evaluation-metrics-for-object-detection/
[14] https://qiita.com/yu4u/items/7e93c454c9410b4b54279f21
[15] Confusion Matrix in Object Detection with TensorFlow https://towardsdatascience.com/confusion-matrix-in-object-detection-with-tensorflow-b9640a927285
[16] https://github.com/spvino/tf_object_detection_cm