International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 12 | Dec 2021 www.irjet.net p-ISSN: 2395-0072

Real Time Object Detection and Tracking using Mask R-CNN

Meghana H S*, Kiran K*, P Deep Shenoy*, and Venugopal K R+

*Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bengaluru, India
+Bangalore University, Bengaluru, India

Abstract: *Object recognition and tracking are critical for a wide range of applications, including security, monitoring, man-machine communication, automotive assistance systems, traffic management, and others. The goal of all of these applications is to identify the target object and locate it at regular time intervals. While individuals must be identified and traced in fatal accidents in safety applications to protect them from harm from automobiles, they are identified in the detection system to assess their movement pattern for conformance to a specified standard for social safety and access [1]. Detecting objects is necessary to start the tracking. Using temporal data obtained from a succession of photos, such as estimating inter-frame variance, training a static backdrop scene model and correlating it to a local scene, or discovering high motion places, is a typical strategy for recognizing the moving objects. The purpose of this work is to create a compelling framework for instance divisions. The vision group has rapidly enhanced content recognition and semantic division performance in a short period of time. Reliable standard methods, such as the Fast/Faster RCNN and Fully Convolutionl Network (FCN) object identification and semantic division frameworks, have largely governed these improvements.*

Keywords: Fully Convolutionl Network (FCN), Image Processing, Object Detection, Object Tracking, Region Based Convolutionl Neural Network (R-CNN), Tensor Flow

1. INTRODUCTION

The main objectives of applications such as safety, monitoring, human-robot communication, Automation driving systems, traffic management, smart robots, clinical logic, and many others are to detect and track objects. The goal of all these technologies is to transfer human visual processing abilities to computers and robots. Despite the noise density, the large size of visual data, the perception of visual data to modify illumination conditions, the difficulty of the input image and movement, non-rigidity of the object, narrowing, unexpected movement, actual specifications, and several other concerns, visual object recognition and tracking remains a difficult problem. Object recognition and object tracking are becoming most prominent study areas in the machine learning and artificial intelligence communities in this regard. Many studies have been undertaken in recent days, in both hardware and software domains, either to suggest a novel remedy of image classification and object tracking or to enhance existing methods. Indeed, advances in sensor technology, as well as superior computational approaches, have opened up new avenues in the field of machine learning.

Object detection and object tracking are two of the most common and difficult tasks that a monitoring system must complete in order to identify relevant activities and hazardous actions, as well as autonomously caption and extract video information. An object could be a body, a skull, a line of people or a product on an assembly line, according to the business analytics concept. The key trends are introduced to provide a classification of common methodologies with the aim of enabling the combination of object detection and tracking for even more efficient enterprise video surveillance [2]. Motion-based recognition and appearance-based object recognition are the two types of object recognition techniques. The object's speed, momentum, orientation, and trajectory are included in motion-based approaches, while appearance-based processes incorporate attributes such as color, border, form, size, and any other immobile characteristics. Tracking is defined as the process of identifying an object's direction as the object is moving along a scene. Due to the extreme growing need for computerized monitoring system frameworks and the advancement of high-powered computer systems, object detection and object tracking concepts of machine learning are gaining prominence, and are widely used in applications such as artificial intelligence monitoring, human-machine communication, motion-based detection, autonomous driving, traffic management, video encoding etc.

2. LITERATURE SURVEY

Fatih Porikli et. al., [2] proposed a method for learning different perspectives of an object and then use them during tracking. Furthermore, a tracker that incorporates generic limitations using context data is proposed. The one that is based on the shape and movement of items will actually do better than the one that is based on the form and movement of objects. This data is not exploited. The capacity to learn modeling tools online could substantially expand a tracker's application. Therefore, if available information sources

are particularly preceding and context data, they must be used to fine-tune the tracker to the specific case. A methodical strategy to bring these different sources of data together will yield a basic tracker that could be used successfully in business intelligence tools.

Rakesh Chandra Joshi et. al., [3] proposed a work that organizes and classifies all of the R-CNN approaches. Following the detection of an object, categorization is carried out in order to trace it down later. Object tracking is a technique for locating the target in consecutive video frames. Various obstacles in video make monitoring complicated. Illumination variance, co-ordinates alignment, atmospheric difficulties, tracking object variation, posture variation, interference, motion blur, and so on are some of the criteria that characterise object tracking. For the evaluation and rigorous examination of all conceivable strategies, a large classification is explored, such that each element is addressed.

Several cameras are required for huge monitoring, and all cameras must be integrated to perform multi-view study of a particular object or to associate the field of view of separate cameras. Mist, smog, rainfall, flame, frost, smoke, and other environmental circumstances might cause wrong object recognition, or the system might lose its sturdiness or be unable to detect effectively. For example, with an automobile monitoring system, a vast range of cars are accessible in terms of size, colour, trademark, and form. To identify and track effectively, the system must be taught with various physical aspects of the car.

Mukesh Tiwari et. al., [4] aim at analyzing and examining the prior technique to object tracking and detection employing video sequences through various stages. They also define the difference and provide a novel method for improving object tracking across video frames. Several object recognition, monitoring, and identification approaches, as well as feature descriptors and segmentation methods based on video frames and tracking approaches are assessed in this study. With new concepts, this analysis was adopted to improve object recognition.

Due to frequent changes in object movement and fluctuation in scene size, occlusions, visual modifications, and ego-motion and lighting modifications, Mukesh Tiwari et. al., [4] have emphasized object identification and tracking as one of the most important fields of study. Selection of features, in particular, is critical in object detection. It is used in a variety of real-time activities, such as vehicle observation and video monitoring. To resolve the problem of recognition, monitoring of object movement and visibility is used.

Kumar S. Ray et. al., [5] applied SP Theory of Intelligence to the basic challenge of detecting moving things in video pictures. Gerard J Wolff initially proposed the SP Theory of Intelligence, a foundation for artificial intelligence, in which S stands for Simplicity and P represents Power. They discover and distinguish objects of interest in image sequences with multilevel hierarchy portions and subsections depending on polythetic classifications using the idea of multiple alignments. Identification of objects in a video scene and monitoring those objects in an actual video feed is a difficult task for any visual monitoring system. Authors perform multiple tasks in this work, including extracting correlated data and interpreting actual data lines to achieve optimum noise free coordination for more precise identification. Authors have acquired ideal approach through the use of family resemblance or polythetic aspect, analyzing a scene with high-level feature orientations captured with more rational attention to detail its presence in the actual data.

Mohana et. al., [6] maintain the exclusivity of cutting-edge networks like DarkNet. On a dataset of urban vehicles, effective detection and tracking can be seen. Real-time, exact, specific identifications are provided by the techniques, which are ideal for actual traffic applications. For computer vision systems, object recognition and tracking go together. Object detection is the process of identifying an object or finding a particular case of interest within a collection of suspect images. Object tracking is the process of determining a trajectory or direction that an object follows in multiple frames. A collection of frames is the image obtained from the dataset. In today's technology age, information is the new oil. The cause of effective data has altered expected outcomes with regard to speed and precision. The improvement is seen since the data is processed using two industry phrases: Computer Vision and Artificial Intelligence. For traffic monitoring technologies, these two technologies have enabled key duties such as object detection and tracking.

Shaojian Song et. al., [7] use data augmentation approaches to compensate the dataset's lack of accurate depiction of the discrepancy. The equalisation of ultra low frequency traffic signs enhances prediction performance, according to observational data. Deep-Sort is introduced to the video detection system to achieve video prediction performance. This effectively eliminates false detection and missed detection due to external influences and enhances the performance of the proposed system. In actual traffic settings, autonomous vehicles must identify and track traffic signs, that give significant predictive analytics.

Chandan G, et. al., [8] demonstrated excellent detection and tracking performance on the trained object that could be used in specific circumstances to identify, track, and focus on specific targeted objects in video surveillance. This real-time environment assessment may produce excellent outcomes for any business by allowing security, control, and efficiency. The main goal of the Single Shot Detector method is to identify and monitor numerous objects in an actual video stream. Deep learning has had a significant impact on how the world has adapted to Artificial Intelligence in recent times. Region-based Convolutional Neural Networks and Faster-RCNN are two common object detection techniques. These techniques detect objects quickly and efficiently without loss of performance.

Akhil Addapa et. al., [9] intend to include cutting-edge object detection techniques with the vision of attaining high precision and real-time performance. Many object identification systems face a key difficulty where they rely on other image processing techniques to aid the deep learning-based methodology, which results in poor and sub-optimal efficiency. The authors employ a comprehensive deep learning strategy to mitigate the issue of object recognition in this work from the beginning to the end. Machine learning is the computational model that computer systems use to execute a task without utilising detailed instructions, rather depending on correlations and interpretation. Artificial intelligence is regarded as a subset of it.

Kaiming He et. al., [10] provided an object instance segmentation system that is relatively simple, robust, and generic. The method accurately identifies objects in an image while also producing a high-quality segmentation mask for every instance. Mask R-CNN is an extension of Faster R-CNN that adds a branch for estimating an object mask in addition to the current branch for bounding box recognition.

3. PROPOSED METHODOLOGY

3.1 Image pre-processing

The input image is pre-processed and sent through the network. The Figure 3.1 shows the steps of image pre-processing.

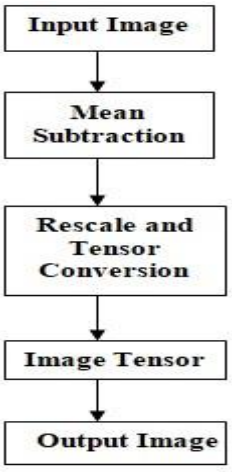

Figure 3.1: Steps for Pre-Processing

- **Mean subtraction**: The mean vector represents the average of all data into pixel values. In addition, test images are removed from the information outline.

- **Re-scale:** The two borders that follow are determined from the target size and most extreme size. The smaller side is scaled to fit the target size, while the bigger side is adjusted to keep the proportion of the maintained tilt.

3.2 Mask R-CNN

Mask R-CNN is a Deep neural organization focused on resolving the AI or machine learning occurrence division problem. Overall, it can perceive a variety of things in an image or video. When given a picture, it responds with boxes, gatherings, and covers that indent the item. The advancement of pixel-level assertion of outlines for articles is known as instance segmentation. Outline for articles is one of the most difficult vision tasks available, and it stands out among similar machine learning problems.

3.3 Faster R-CNN

A Region Proposal Network is a critical interface that provides boundary boxes for prospective items. The following technique, known as rapid R-CNN, uses RoIPool to remove features from each contender box and conducts depiction and bounding box regression. The traits employed by the two steps is presented for quick guessing. For constant segregated evaluations, scrutinizers are incorporated to Faster R-CNN and numerous designs.

3.4 TensorFlow

TensorFlow is an open-source programming library. TensorFlow was created by Google Brain Team

developers for the purposes of guiding machine learning and deep neural network exploration within Google's Machine Intelligence research division.

3.5 Data Flow Diagram

The dataflow diagram provides the graphical portrayal of stream of data through the system. It assists in making the system outline. Figure 3.2 shows the dataflow diagram of Mask R-CNN.

A CNN is utilized by the Region Proposal Network to make the difference of ROI using a lightweight equal classifier. This is accomplished by placing 9 anchor boxes over the image. Objects/no-objects scores are returned by the classifier. Rather than a single distinct one, the RoI Align network generates a few of jumping boxes and rearranges them into a static guess.

The breaking point box hypothesis is modified utilizing the recurrence model after twisted features are treated with completely associated layers to group with SoftMax. Misshaped attributes are then treated with entirely linked layers in order to coordinate with SoftMax, and leaping box speculation is also Enhanced utilizing the backslide method.

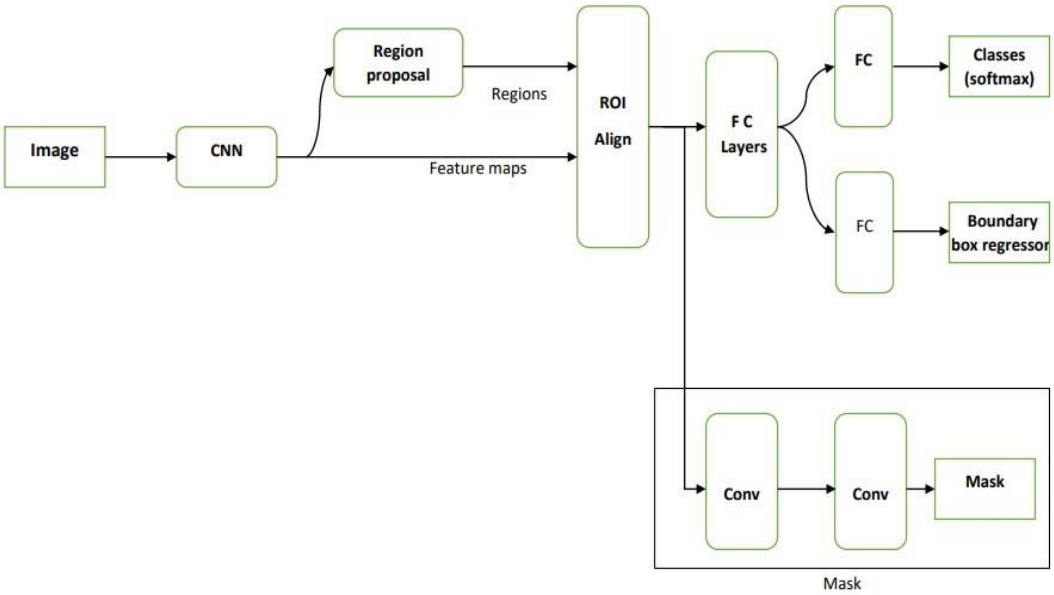

Figure 3.2: Data flow diagram of Mask R-CNN

The Mask classifier is also used to cope with twisted credits, which uses two CNNs for every row to offer a double mask. The Mask Classifier enlists the help of the union to create a mask to every class avoiding putting them in conflict.

3.6 Implementation details

3.6.1 Dataset

The Stanford 2D-3D Dataset is used here. The dataset contains 70,496 RGB photos with associated segmentation images coded in Green and Blue channels. The photos are 1080X1080 pixels, padded to 1088X1088 pixels to keep the tensor forms realistic and divisible by 32. The collection contains 6 zones, 13 object classes, 11 categories of scenes, and 270 scene patterns. and 10 images from a shuffled population of 10,000 images from areas 3 and 5a [Fig 4.2 and 4.3] image of the bounding boxes are considered.

3.6.2 SCIKIT- Learn

Scikit-learn is a built-in package in the Python programming language. It is a sophisticated library package that is mostly used for several machine learning classification techniques. It is a useful tool that is primarily used for data mining and analysis approach.

3.6.3 Scipy

Scipy, Python's reasoning toolkit, is a free software math, science, and design library that is BSD-authorized. The Scipy package uses NumPy to manage the N-dimensional display in an effective and quick approach. Clusters in NumPy must work, thus the Scipy library was designed with that in consideration

3.6.4 NumPy

NumPy is a key Python package for calculating. It creates a multidimensional array object that is superior. It also includes specific equipments for working with these arrays. It contains complicated functions, the robust N-

dimensional array object and also the tools for integrating the codes that are based on Python

3.6.5 OpenCV

Object detection is a helpful technique that describes the algorithm's construction and gives a practical example. For this purpose, OpenCV (Open-Source Computer Vision Library), a free software library for computer vision and image processing applications that is simple to integrate into Python is used.

3.6.6 Anaconda

Anaconda is a Python and R distribution that is free software. Its machine learning and data analytics software that is available for free. NumpPy, Scikit-learn, and pandas are some of the programmers that are associated into the anaconda package. Predictive modeling, deep learning, and data analytics projects all use these built-in packages.

4. RESULTS

This section shows the results of the proposed work which includes accuracy and Confusion matrix values. Architecture of confusion matrix is shown in Table 4.1.

Table 4.1: Confusion matrix architecture

		Predicted class	
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

A confusion table is a matrix utilized to determine the effectiveness of categorization on test data when the true values are known. True positive and true negative scores, which are shown in green, are accurately predicted classes. The purpose is to minimize the amount of false positive and negative readings (highlighted in red).

Screenshots of Mask R-CNN

In figure 4 .1, Mask R-CNN model of object detection and tracking system detects the human being as a person with accuracy 1.00 [100%].

Person 1

Figure 4.1: Person 1

Person 2

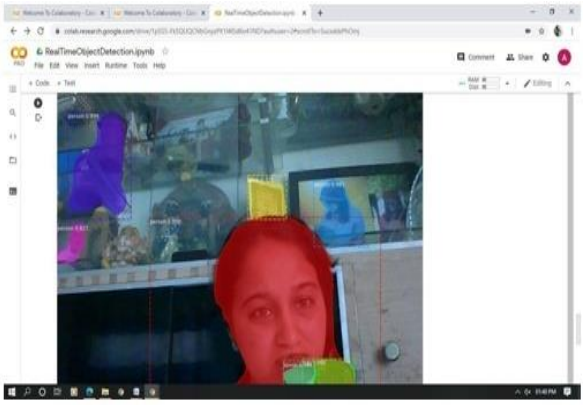

Figure 4. 2: Person 2

In figure 4.2, it detects the person and some photos of the human being as person and accuracy of the Person is 1.00 [100%] and that of person image is 0.994

Person 3

Figure 4. 3: Person 3

Person 4

Figure 4.4: Person 4

In figures 4.3 and 4.4, it detects various things like photos of the God and person but sometimes it gives false result, like in the Figure 4.4, the flower is detected as Umbrella and accuracy of the person is 1.00 [100%] and umbrella is 0.872.

Person 5:

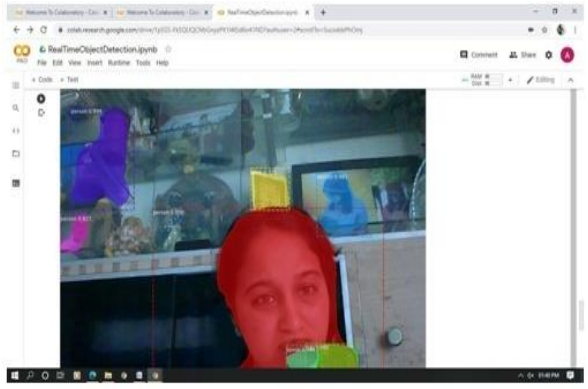

Figure 4.5: Person 5

In Figure 4.5, it detects the person, and some photos of the human being as person and accuracy of the Person is 1.00 [100%] and that of person image is 0.994.

Person 6

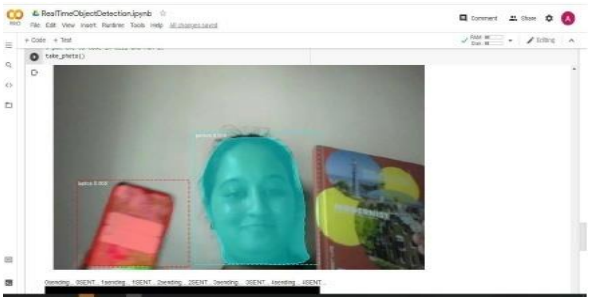

Figure 4. 6: Person 6

In the figure 4.6, it detects the person, book and cell phone. Accuracy of the book is 0.974, person is 1.00 [100%], and the accuracy of the cell phone is 0.868.

4.2 Trained Dataset values of the result

Table 4.2: Iteration 1

Y_act	Y_pred	True	False	Accuracy
2	2	True	-	1.00
2	60	-	False	0.78

In table 4.2, the person's real value in the data set is 2 and the predicted value in iteration 1 is also 2, indicating that the person's accuracy is 1[100%] and the result is true.

Table 4.3: Iteration 2

Y_act	Y_pred	True	False	Accuracy
2	2	True	-	1.00
2	88	-	False	0.871
1	2	True	-	0.974
44	3	-	False	0.705

The actual value of person is 2 in table 4.3, so it is true, and sometimes image of the person or teddy bear and spoon it taken as person or some other training data set with the accuracy of 0.871 for teddy bear, and 0.971 for spoon, so it is false.

Table 4.4: Iteration 3

Y_act	Y_pred	True	False	Accuracy
2	2	True	-	1.00
2	60	-	False	0.78

We take the real values of person and cell phone in table 4.4. Since the anticipated and actual values of the person are the same, the genuine result is obtained with an accuracy of one. However, if the person in the photograph is identified as a laptop, it returns a false result with an accuracy of 0.78.

Table 4.5: Iteration 4

Y_act	Y_pred	True	False	Accuracy
7	1	True	-	0.999
78	78	True	-	0.834
78	75	-	False	0.723

Table 4.5 shows the actual result of a person riding a bicycle, which matches the predicted data value of 1 and yields the true result with accuracy of 0.999 that is equal to 1, and the false result of a microwave with an accuracy of 0.723 in the same iteration.

Table 4.6: Iteration 5

Y_act	Y_pred	True	False	Accuracy
2	1	True	-	0.99
1	84	-	False	0.782
1	28	-	False	0.758
1	1	-	False	0.940

Table 4.7: Iteration 6

Y_act	Y_pred	True	False	Accuracy
27	27	True	-	0.813
84	84	True	-	0.865
	2	True	-	0.998

Tables 4. 6 and 4.7 shows the genuine results of the bag, book, and person, with the backpack's accuracy of 0.813, the book's accuracy of 0.865, and the person's accuracy of 1.

4.3 Confusion matrix of the result

Figure 4.3: Confusion matrix

The confusion matrix is a cross table that has class detection counts in the columns and class ground truth counts in the rows as shown in the fig 4.3. True positives, or instances that have been accurately recognised and categorised, are represented by the diagonal cells of each associated class. Counts of false positives are at the bottom of each column, while counts of false negatives are at the top of each row.

5. CONCLUSIONS

The goal of this work is to create an optimized Mask RCNN model employing Tensor Flow's CPU variant to identify and monitor real-time objects utilizing a real-time camera. This is accomplished by using many layers in a deep neural network to learn properties of varying sizes, such as anchors and RoI Align, rather than treating layers as a black box, as is the case with traditional CNN-based object detection techniques. RCNN results in accuracy ranging 70-100%, 0.8888889 precision and 0.5454545 F1- score. The findings show that this technique has a high degree of precision in object detection caught in real-time motion.

6. REFERENCES

[1] "Real Time Object Recognition and Tracking Using 2D/3D Images," 2010.

[2] Fatih Porikli and Alper Yilmaz, "Object Detection and Tracking," Video Analytics for Business Intelligence, pp. 3-41, 2012.

[3] Rakesh Chandra Joshi, Mayank Joshi, Adithy Gaurav Singh and Sanjay Mathur, "Object Detection, Classification and Tracking Methods for Video Surveillance: A Review," International Conference on

Computing Communication and Automation, pp. 1-7, 2019.

[4] Mukesh Tiwari and Rakesh Singhai, "A Review of Detection and Tracking of Object from Image and Video Sequences," International Journal of Computational Intelligence Research, pp. 745-765, Vol.13,2017

[5] Kumar S. Ray, Sayandip Dutta and Anit Chakraborty, "Detection, Recognition and Tracking of Moving Objects from Real-time Video via SP Theory of Intelligence and Species Inspired PSO," pp. 1-16.

[6] Mohana and HV Ravish Aradhya, "Object Detection and Tracking using Deep Learning and Artificial Intelligence for Video Surveillance Applications," International Journal of Advanced Computer Science and Applications, pp. 517-530, Vol. 10, 2019.

[7] Shaojian Song, Yuanchao Li, Qingbao Huangand Gang Li, "A New Real-Time Detection and Tracking Method in Videos for Small Target Traffic Signs," Applied sciences, pp. 1-16, 2021.

[8] He, K., Gkioxari, G., Dollár, 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.: 'FasterR-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,

[10]https://github.com/tensorflow/models/blob/maste research/object_detection/g3doc/tf1_detection_zo .md

[11]Haruhiro Fujita, Masatoshi Itagaki, KentaIchikawa, Yew Kuwan Hooi, Kazuyoshi Kawahara and Aliza Sarlan, Fine-tuned Surface Object Detection Applying Pre-trained Mask R CNN Models, International Conference of Computational Inteligence 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/a introduction-toevaluation-metrics-for-object detection/

[14] https://qiita.com/yu4u/items/7e93c454c9410c42 #fn21