AUTOMOBILE BODY DAMAGE DETECTION USING COMPUTER VISION

Amit Birajdar
Computer Engineering
MPSTME , NMIMS
Mumbai , India
amitbirajdar7@gmail.com

Harsh Agarwal
Computer Engineering
MPSTME , NMIMS
Mumbai , India
harsh30199@gmail.com

Manan Bolia
Computer Engineering
MPSTME , NMIMS
Mumbai , India
mananbolia14@gmail.com

Vedang Gupte
Computer Engineering
MPSTME , NMIMS
Mumbai , India
vedanggupte@gmail.com

***Abstract*— Dent detection is an emerging application involving multiple disciplines like artificial intelligence, machine learning and image processing. Vehicular accidents are at an all-time high and are only projected to increase with a rise in the number of vehicles on road. The traditional process of dent detection involves human inspection and judgement to identify structural damage. Assessment of the magnitude of damage also involves expert analysis. This is a time consuming process involving multiple middlemen. There are various methods to detect dents without involving any third-party and in a relatively short time. In this paper we review four such methods of dent detection. We start with discussing how dents are detected at the factory level. We then move on to three-dimensional surface analysis to detect dents without any prior knowledge of the part specification. We then observe two general object detection methods- YOLO and Mask RCNN. Their implementation and limitations are also discussed.**

Keywords*— *You Only Look Once, Mask-RCNN, Damage Detection, Quantization , 3D Surface Analysis

I. INTRODUCTION

Dent detection works on the basic principles of general object detection. Detection starts at the factory-level itself where defective parts have to be identified and rejected. At this stage, the dimensions of each part are known so the implementation is relatively simpler. Cameras at various angles capture images and a neural network is trained to identify intolerances. However, this method fails if the exact dimensions of the parts are not known. The other three methods work without explicit knowledge of original dimensions. Three-dimensional analysis focuses on using image processing techniques to identify uneven surfaces which are potential dents. Areas of the image are segmented until the parts are identified as either dents, dings or undamaged area. YOLO method uses regression to detect objects. In this, we won't select the interested regions from the image. Instead, we predict the classes and bounding boxes of the whole image at a single run of the algorithm and detect multiple objects using a single neural network. YOLO algorithm is fast as compared to other classification algorithms. In real time this algorithm processes 45 frames per second. YOLO algorithm makes localization errors but predicts less false positives in the background. Mask R-CNN, extends Faster R-CNN by adding a branch for predicting segmentation masks on each Region of Interest (RoI), in parallel with the existing branch for classification and bounding box regression. The mask branch is a small Fully Convolutional Network (FCN) applied to each RoI, predicting a segmentation mask in a pixel-to pixel manner. Mask R-CNN is simple to implement and train given the Faster R-CNN framework, which facilitates a wide range of flexible architecture designs. Additionally, the mask branch only adds a small computational overhead, enabling a fast system and rapid experimentation.

II. DETECTION AT FACTORY LEVEL

This method detects small dings in the sheet metal of cars where the panels have not been painted or the paint remains undisturbed. The imaging apparatus consists of two cameras and a projector. Neural networks are used for error detection. Two requirements are imposed- the optical measurement system needs to be accurate , and the master(undamaged) and test(possibly damaged) panels can have minute differences at a threshold which cannot be detected easily if the surface area is large. The two cameras are positioned at different angles and a projector is used to focus light in different patterns(usually 12) on the test piece to test surface aberrations. Images are clicked in rapid succession to capture the intensity levels at different points. Initially in the training phase, measurement values of several master panels are captured at different angles to create the data for the neural network .The different measurements of a panel are then mixed with measurements of another so the a large data set can be generated with the required tolerance .The calculation of weights for the network then follows. In the recall phase, the test piece is measured at different angles, and compared with the data of the neural network. The difference is then calculated. The network is a modified associative memory, it has 2 active layers. The z coordinate of the data gathered is used for calculations. A matrix zi is the matrix with z coordinates of i^{th} set, matrix Z is the matrix containing matrices of all sets i to n. The weights are calculated using Z=USV where S is a diagonal matrix with nonnegative diagonal elements (singular values of 2) in decreasing order. U contains the eigenvectors of the covariance matrix ZZ'. The first m columns of U represent the weight matrix W of the neural network. The accuracy of the algorithm is then tested using car doors. 9 out of 10 doors serve as master pieces. 45 data sets are then built. The training time is 5 hours for each region. The 10th door is flawed and serves as the test case. The error is correctly detected with x and y coordinate of defect as output. Another flawless piece is tested which was not part of the training data. It is passed as error free by the algorithm as intended. 1 second was required for the calculations in both cases.

III. 3D Surface Analysis

This paper focuses on the analysis of 3D surfaces and automatic detection of deformations. One important contribution of this work comes from the imposed requirement that this system must be able to detect deformations without knowledge of the ideal shape of the part. This means that no model needs to be stored in memory of the ideal faultless part. Multiple techniques are discussed here that achieve this objective. Some basic parameters about the size of deformations needs to be decided first. The deformation detection system consists of a surface shape analysis phase to extract areas of interest, a segmentation phase to group areas containing pieces of deformations together into segments, and a classification phase to determine which segments contain deformations and which contain design features. The paper first discusses how the images can be captured. Laser scanners, stereoscopic vision systems and Xbox Kinect are also considered for getting 3D models. The paper then talks about surface analysis to locate the defects in question. Since no prior knowledge of the object is present , Well-known edge detectors, such as the Sobel and Canny operators, can highlight the areas that belong to features. K means algorithm or Unseeded region growing can also be used for this process. The proposed system takes a 3D mesh as an input, and outputs the sections of the mesh which are deformations of interest along with whether they are a ding or a dent. The proposed system contains 3 major components. The surface shape analysis component is tasked with dividing the 3D mesh into sections and analyzing each one for the magnitude of the deformation contained in that section. The segmentation component combines sections from the surface shape analysis which seemingly belong to the same deformation. The classification component classifies each segment from the segmentation as either a ding or dent, and removes segments which do not meet the criteria of being a deformation of interest, such as vehicle design features and acquisition noise. For surface shape analysis, the octree method can be used,which divides the mesh into cubes, which can be further subdivided to give an estimation of where the deformations can exist. An evolution of this is triangle based analysis instead of point based analysis of surface normal. The surface normal of triangles can be compared to standard deviation to detect presence of deformations .Moreover, unequal weight can be given to the normal based on surface area of triangles to minimize the effect of small noisy areas. Since this is dependent on the standard deviation at a particular resolution, multiresolution analysis is important to calculate an aggregate standard deviation which will give good results at any resolution. The paper has compared local and aggregate SD to show how taking an aggregate is better and more reliable even when threshold changes. In the second stage, the algorithm must segment the data into areas with deformations and areas without deformation. The octree is used to segment the object. Sufficient resolution is necessary to distinguish between areas, however it should be low enough so that noise does not affect the output. Once the object is divided into feature and non-feature areas, the segmented areas containing features must be recombined based on adjacency using their coordinates. Finally comes the classification phase. It classifies the detected areas into dings or dents, and also removes some false positives that may remain from the previous phases . A least-squares plane-of-best-fit fitted to the 3D points contained in a segment, specifically the boundary points, is used to determine the orientation of the shape represented by a given segment. A descriptor, called the point-count descriptor, uses the number of points that share a similar positional relationship to the plane-of-best-fit in estimating the direction of variation of the surface contained in the segment. If a majority of the points contained in the segment are above the plane-of-best-fit, that is, in the direction of the normal vector, the deformation is classified as a ding. If a majority of the points are below the plane-of-best-fit, that is in the opposite direction to the normal vector, the deformation is classified as a dent. The percentage of the points that are above the plane-of-best-fit in the case of a ding, or below the plane-of-best-fit in the case of a dent, provides the certainty measure on the classification.

IV. You Only Look Once

You Only Look Once (YOLO) uses regression to detect objects. In this algorithm, classes and bounding boxes of the whole image are predicted at a single run of the algorithm and multiple objects are detected using a single neural network. YOLO algorithm is fast as compared to other classification algorithms.

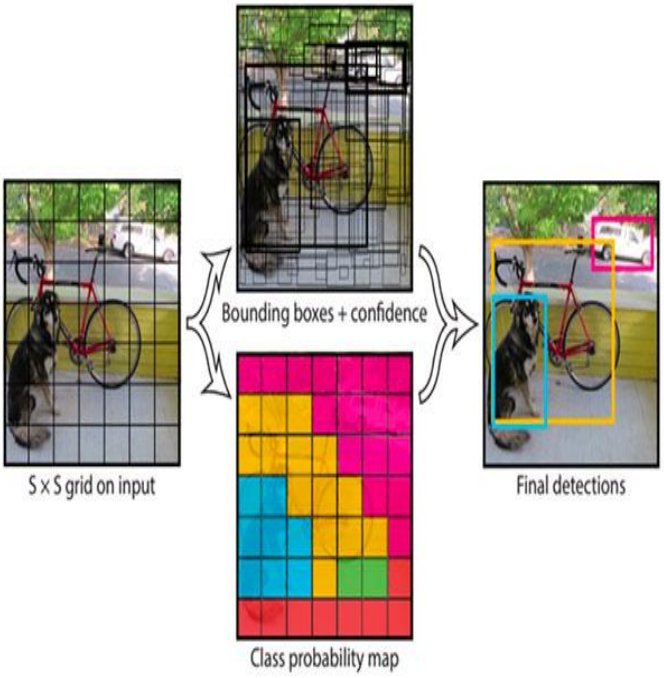

Figure 1: [5] Working of YOLO

First, an image is taken and YOLO algorithm is applied. The image can be divided into any number grids, depending on the complexity of the image. Once the image is divided,

each grid undergoes classification and localization of the object. The objectness or the confidence score of each grid is found. If there is no proper object found in the grid, then the objectness and bounding box value of the grid will be zero or if there is an object in the grid then the objectness will be 1 and the bounding box value will be the corresponding bounding values of the found object.

YOLO algorithm is used for predicting the accurate bounding boxes from the image. The image is divided into S x S grids by predicting the bounding boxes for each grid and class probabilities. Both image classification and object localization techniques are applied for each grid of the image and each grid is assigned with a label. Then the algorithm checks each grid separately and marks the label which has an object in it and also marks its bounding boxes. The labels of the grid without object are marked as zero.

If two or more grids contain the same object then the center point of the object is found and the grid which has that point is taken. For this, to get the accurate detection of the object two methods can be used. They are Intersection over Union (IOU) and Non-Max Suppression. In IoU, it will take the actual and predicted bounding box value and calculate the IoU of two boxes by using the formulae, IoU = Area of Intersection / Area of Union. If the value of IoU is more than or equal to the set threshold value (0.5 in this case) then it's a good prediction. The threshold value is just an assuming value. The authors also recommend taking greater threshold value to increase the accuracy or for better prediction of the object. The other method is Non-max suppression, in this, the high probability boxes are taken and the boxes with high IoU are suppressed. This is repeated until a box is selected and that box is considered as the bounding box for that object.

By using Bounding boxes for object detection, only one object can be identified by a grid. So, for detecting more than one object an Anchor box is used. Anchor boxes are a set of predefined bounding boxes of a certain height and width. These boxes are defined to capture the scale and aspect ratio of specific object classes you want to detect and are typically chosen based on object sizes in your training datasets. During detection, the predefined anchor boxes are tiled across the image. The network predicts the probability and other attributes, such as background, intersection over union (IoU) and offsets for every tiled anchor box. The predictions are used to refine each individual anchor box. Multiple anchor boxes can be defined, each for a different object size.

The network does not directly predict bounding boxes, but rather predicts the probabilities and refinements that correspond to the tiled anchor boxes. The network returns a unique set of predictions for every anchor box defined. The final feature map represents object detections for each class. The use of anchor boxes enables a network to detect multiple objects, objects of different scales, and overlapping objects.

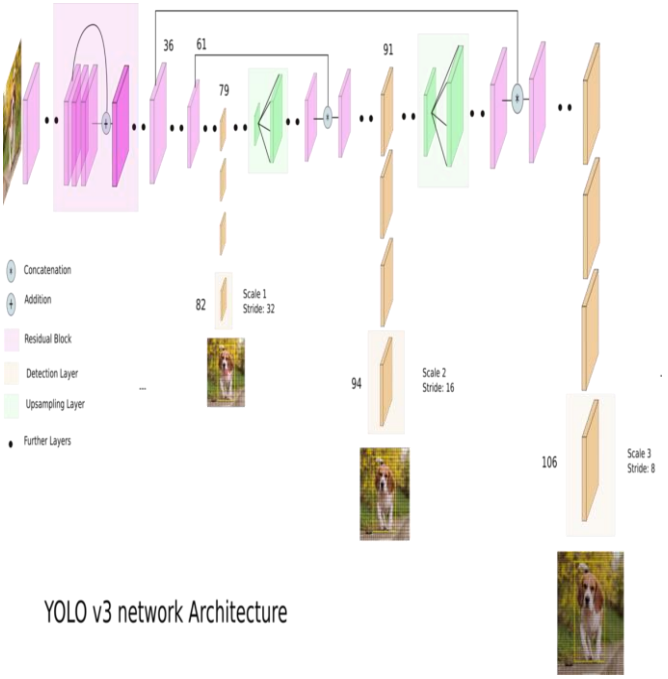

Figure 2: [6] YOLO network architecture

V. MASK-REGIONAL CNN

Mask R-CNN has three outputs for each candidate object, a class label, bounding-box offset, and a branch that outputs the object mask. Mask R-CNN consists of two stages. The first stage, called a Region Proposal Network (RPN), proposes candidate object bounding boxes. In the second stage, in parallel to predicting the class and box offset, Mask R-CNN also outputs a binary mask for each RoI. This is in contrast to most recent systems, where classification depends on mask predictions.

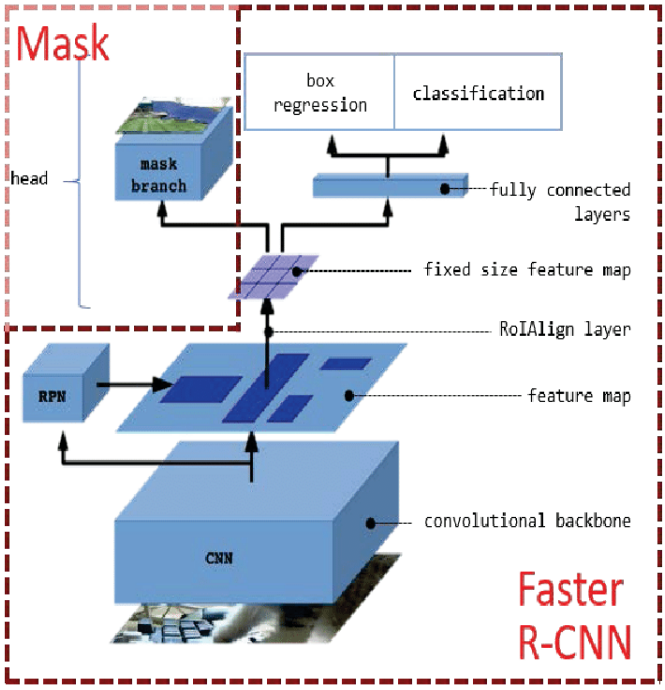

Figure 3: [7] Mask R-CNN architecture

Formally, during training, a multi-task loss is defined on each sampled RoI as L = Lcls + Lbox + Lmask. The classification loss Lcls and bounding-box loss Lbox are identical as those defined in Faster RCNN. The mask branch has a Km^2 - dimensional output for each RoI, which encodes K binary masks of resolution m × m. A mask encodes an input object's spatial layout. Thus, unlike class labels or box offsets that are inevitably collapsed into short output vectors by fully-connected (fc) layers, extracting the spatial structure of masks can be addressed naturally by the pixel-to-pixel correspondence provided by convolutions. Specifically, an m × m mask is predicted from each RoI using an FCN. This allows each layer in the mask branch to maintain the explicit m × m object spatial layout without collapsing it into a vector representation that lacks spatial dimensions of the K classes. To this a per-pixel sigmoid is applied, and a va;ue of Lmask is defined as the average binary cross-entropy loss. For an RoI associated with ground-truth class k, Lmask is only defined on the k^{th} mask(other mask outputs do not contribute to the loss).

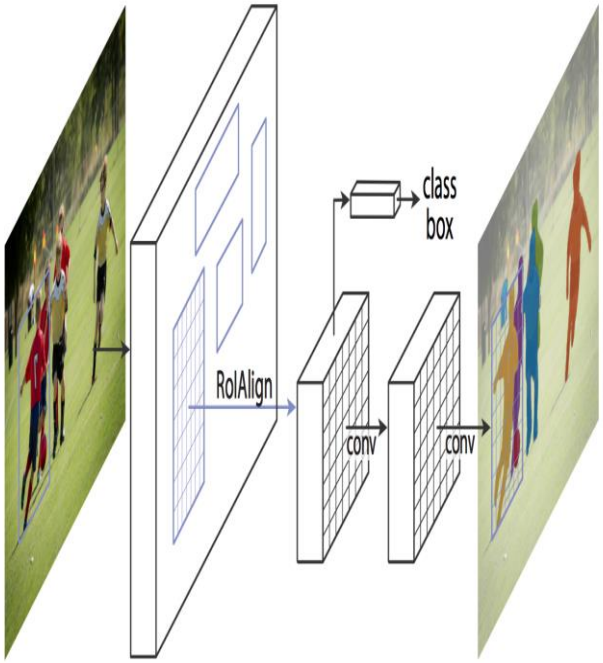

Figure 4: [9] Working of Mask R-CNN

In first stage , a light weight neural network called Region Pooling Network (RPN) scans all Feature Pyramid Network (FPN) top-bottom pathway called feature map and predicts regions which may contain objects. To represent the location of these regions in the input image , anchors are used. " [8] Anchors are a set of boxes with predefined locations and scales relative to images." According to some Intersection over Union (IoU) values , bounding boxes and ground-truth classes are assigned to each anchor. Also anchors of different scales are bind at different levels of feature map , RPN uses anchors to propose where in the image there is an object and what its bounding box should be.

In the second stage , another part of neural network is used to predict the classes , bounding boxes , pixel-level mask for each object. This neural network takes proposed object regions from previous neural network RPN and then scans those regions and predicts object classes , bounding box and mask for each object above an confidence value. It looks similar to previous stage but it uses a different method called RoIPool. RoIPool is a standard operation for extracting a small feature map from each RoI. RoIPool first quantizes a floating-number RoI to the discrete granularity of the feature map, this quantized RoI is then subdivided into spatial bins which are themselves quantized, and finally feature values covered by each bin are aggregated (usually by max pooling). Quantization is performed, e.g., on a continuous coordinate x by computing [x/16], where 16 is a feature map stride and [·] is rounding; likewise, quantization is performed when dividing into bins. These quantization introduce misalignments between the RoI and the extracted features.

VI. SUMMARY TABLE

Reference No.	Aim	Inferences
1	The effect of different CNN configurations on textured-surface defect segmentation and detection performance	CNN architectures were compared for detection and segmentation of textured surfaces. It is not useful for dent detection as it is very complex and segments handles and other protrusions as well which is not required
2	Implement "You only look once" and define set up and parameters in real-time	It is useful for fast detection of objects in videos but lacks in accuracy and finding small objects
3	Using Faster RCNN for object detection	Faster RCNN uses a selective search algorithm to identify region proposals. The regions are then reshaped using RoI pooling layer, which is astronomically faster than earlier approaches
4	Using Mask RCNN for object detection	It is useful as it provides instance segmentation with very good accuracy and fast detection so it is very useful

VII. EXISTING INDUSTRY TECHNOLOGIES

Reference No.	Aim	Inferences
1	Dent detection Using well-known optical flow algorithms and the deflectometry principle	Fast enough to meet assembly line industry standards. Resource intensive and also requires specialised equipment. Cannot detect scratches
2	Automatic dent detection for unpainted panels	Fast and accurate but requires prior knowledge of exact dimensions of parts. Fails if paint is damaged and cannot detect scratches
3	Classification of dents and scratches by feature extraction and a support vector machine algorithm	Almost 100% accuracy and already used in the automotive industry. Can differentiate between dents and scratches, and also estimate size of deformations. However it requires specialized , expensive equipment and cannot work using smartphone cameras
4	Dent detection without prior knowledge of parts	Reasonably good accuracy. Can work on a variety of cars. Does not work well if the design of car panels naturally includes many edges or protrusions and detects false positives ,and cannot differentiate between dents and scratches

VIII. CONCLUSION

At factory level[1], the focus is on how to detect errors in cases where paint damage cannot be used to differentiate between faulty and flawless panels. The training time is adequately fast and the calculations and error detection is fast(1 second or so) and accurate within the specifications dictated by the author(20 micrometre).However this algorithm cannot differentiate between small and big dents. Thus more work needs to be done to gauge the size of the dings as this is an important parameter to calculate insurance cost. It also needs information about the exact measurements of the ideal part so a lot of training is required for different panels. Moreover, the images of the test pieces have to be captured in the exact way as the reference panels, hence it can only be used if the tester also has access to the same setup, otherwise the algorithm fails.

Since prior knowledge and measurements of parts is not always know, it is important to dings and dents without having any prior knowledge about the shape of the panel .Thus 3D surface analysis [2] can be used on a lot of different models ,without any specific training required. It has a 3 stage process which can also classify regions as dings or dents with a calculated level of certainty. The paper states how to get reasonably good results in all three stages. However, often some creases or shapes which are part of the aesthetic can get falsely identified as dents. Door handles can get identified as dings as well.

YOLO is particularly fast , which makes it good for detection in videos, but it lacks in accuracy for small objects which make it not so viable for our application as there is possibility of small damages like dings and scratches. Mask Regional-Convolution Neural Network (Mask RCNN) is the preferred choice as it provides instance segmentation with good speed and accuracy than other object detection algorithms. Mask RCNN also detects small objects successfully, which is a major advantage over other techniques. Although Faster and Fast RCNN have greater speed they lack in accuracy which is compensated by Mask RCNN.

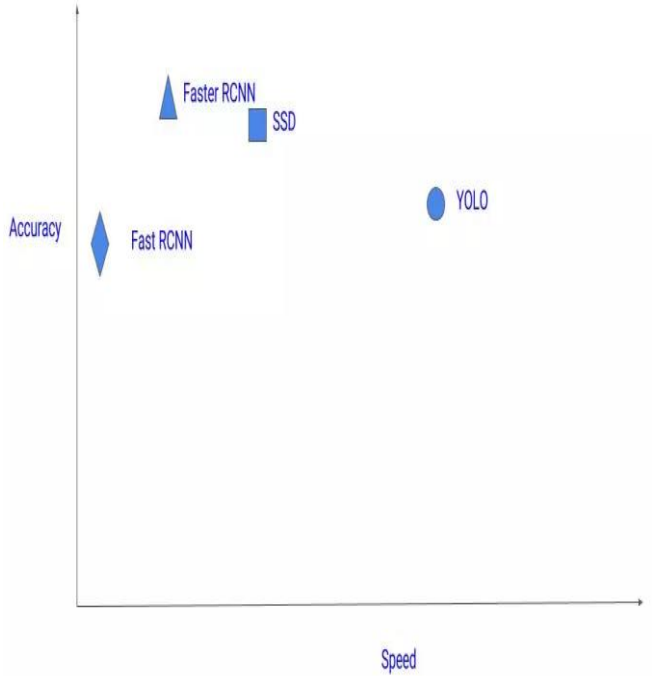

Figure 5: [10] Comparison between different networks

IX. REFERENCES

[1] T. Lilienblum, P. Albrecht, R. Calow and B. Michaelis, "Dent detection in car bodies," Proceedings 15th International Conference on Pattern Recognition. ICPR-2000, Barcelona, Spain, 2000, pp. 775-778 vol.4.

[2] Yogeswaran A, Payeur P. “3D Surface analysis for automated detection of deformations on automotive body panels.” New Advances in

Vehicular Technology and Automotive Engineering. 2012 Aug 1:978-53.

[3] Geethapriya. S, N. Duraimurugan, S.P. Chokkalingam “Real-Time Object Detection with Yolo”,International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8, Issue-3S, February 2019

[4] “He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. "Mask r-cnn." In Proceedings of the IEEE international conference on computer vision, pp. 2961-2969. 2017.

[5] https://pyimagesearch.com/wp-content/uploads/2018/11/yolo_design.jpg

[6] https://towardsdatascience.com/yolo-v3-object-detection-53fb7d3bfe6b

[7] https://www.researchgate.net/profile/Lukasz_Bienias/publication/337795870/figure/fig2/AS:834563236429826@1575986789511/The-structure-of-the-Mask-R-CNN-architecture.png

[8] https://medium.com/@alittlepain833/simple-understanding-of-mask-rcnn-134b5b330e95

[9] https://pyimagesearch.com/wp-content/uploads/2018/11/mask_rcnn_segmentation_types.jpg

[10] https://cv-tricks.com/wp-content/uploads/2017/12/Various-detectors-2.jpg