# Object Detection of Furniture and Home Goods Using Advanced Computer Vision

Narayana Darapaneni
*Director - AIML*
*Great Learning/Northwestern University*
Illinois , USA
darapaneni@gmail.com

Sunilkumar C M
*Student - PGPAIML*
*Great Learning*
Bangalore, India
sunilcmnitk@gmail.com

Mukul Paroha
*Student - PGPAIML*
*Great Learning*
Bangalore, India
mparoha25@gmail.com

Anwesh Reddy Paduri
*Senior Data Scientist*
*Great Learning*
Hyderabad, India
anwesh@greatlearning.in

Rohit George Mathew
*Student - PGPAIML*
*Great Learning*
Bangalore, India
rohitspeak@gmail.com

Namith Maroli
*Student - PGPAIML*
*Great Learning*
Bangalore, India
namith.m@gmail.com

Rohit Eknath Sawant
*Mentor - PGPAIML*
*Great Learning*
Bangalore, India
Rohit.sawant@gmail.com

2022 Interdisciplinary Research in Technology and Management (IRTM) | 978-1-6654-7886-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/IRTM54583.2022.9791508

***Abstract*—Object Detection Technology has been a subject to much research and development due to increasing use of images and videos as data sources and their huge number of applications. Traditional models for object detection had limitations in training and did not use transfer learning for their benefit. With the evolution of deep learning and Neural networks, newer and powerful tools have made way to achieve Object Detection in real-time, with the added advantage of transfer learning and detection of multiple instances of different classes of interest in the given image context. The proposed system is an Object Detection model based on the Single Shot Detector (SSD) algorithm trained with MobileNetV2 feature extraction that can be utilized and integrated in e-commerce, hospitality industry, security and surveillance, real estate, self-driving cars and floor inventory management.**

***Keywords—object detection, furniture, home goods, Advanced Computer Vision (ACV), Single Shot Detector (SSD), MobileNetV2***

## I. Introduction

As a technology, Computer Vision has diverse subfields. Image Classification[15] is an area of computer science that studies how computer algorithms classify or assign class to real-world object instances in images and videos. Another area is Object Detection, which is associated with the process of locating and identifying object instances of specific classes in a given image that contains instances from several classes. An image bound with boxes around the detected object and the name of the object is the output of an object detection algorithm.

The objectives of object classification and localization are combined in object detection. Computer vision is a multidisciplinary branch of artificial intelligence that attempts to imitate human vision's strong capabilities. With successful applications in healthcare, security, transportation, retail, finance, agriculture, and more, the Advanced Computer Vision (ACV)[17] technologies are transforming entire industries and business operations today. Since Convolution Neural Networks (CNN) began surpassing humans in specialized picture identification tasks, computer vision research has moved at a dizzying rate. In the 1980s, the basic architecture of Convolution Neural Networks (CNNs) was devised. Yann LeCun improved on the original design in 1989 by training models to recognize handwritten numbers using back propagation. Computer vision works through visual recognition techniques like Image classification, object detection, Image segmentation, object tracking, optical character recognition, image captioning, etc.

## II. Related Work

### *A. Single Shot MultiBox Detector*

The SSD method uses a feed-forward convolutional network to build a fixed-size collection of bounding boxes and scores for the occurrence of object class instances in those boxes, followed by a non-maximum suppression step to generate final detections. The early network layers are based on the base network, which is a standard architecture for high-quality image categorization.

### *B. MultiBox*

Szegedy's work on MultiBox [9], a method for rapid class-agnostic bounding box coordinate proposals, influenced SSD's bounding box regression methodology. In the researches done on MultiBox, an Inception [25]-style convolutional network is used . The 1x1 convolutions shown below aid in dimensionality reduction by reducing the number of dimensions.

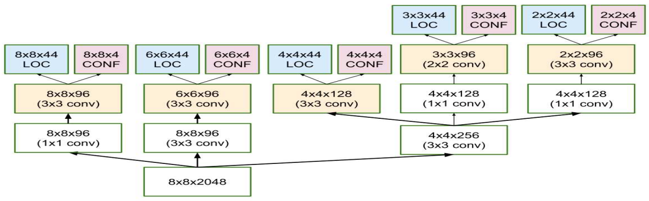

Fig. 1. The 1x1 convolutions

MultiBox's loss function also combined two critical components that made their way into SSD:

978-1-6654-7886-1/22/$31.00 ©2022 IEEE

**Confidence Loss:** Metric indicates how confident the network is in the objectness of the computed bounding box. This loss is calculated using categorical cross-entropy [11].

**Location Loss:** Metric that evaluates how far the network's projected bounding boxes differ from the training set's ground truth bounding boxes. Here, L2-Norm [18] is employed.

### C. MultiBox Priors And IoU

In MultiBox [9] priors, which are pre-computed, fixed size bounding boxes that closely match the distribution of the original ground truth boxes. Those priors are selected in such a way that their Intersection over Union ratio [24] is greater than 0.5. From the image below, an IoU of 0.5 is still not good enough but provides a strong starting point for the bounding box regression.

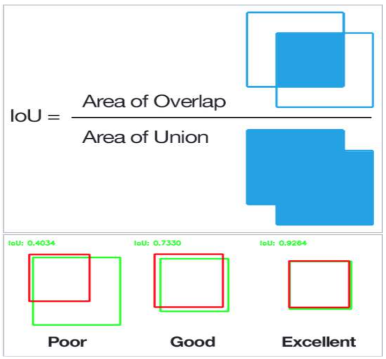

Fig. 2. IoU

### D. MobileNet as a Feature Extractor

1. MobileNet offers a simplified design to generate light weight deep neural networks using depth wise separable convolutions.
2. Adjusts latency and accuracy utilizing two global hyper parameters[19].
3. When compared to other popular models for ImageNet classification, MobileNet outperforms them [6].

## III. Proposed System

### A. Data and Findings

The Input Data for Object Detection method is a digital snapshot of the environment that comprises instances of real-world objects that must be identified. For example, photographs of living rooms, bedrooms, kitchens, and other areas where furniture and other household items can be found.

The intention was to use conventional photographs from Google Images; however, our research indicated that these images would not deliver the intended outcomes since the images were highly irregular in sizes, and the degree of contrast and quality of the images were not uniform. It was decided to use one of the existing curated/standard image sets. With almost 600 item types, Open Images [14] is one such image dataset.

### B. Data Pre-Processing

Data Pre-Processing [22] has proved to be a critical step in achieving optimal performance from the Object detection models. The Open Images [14] Dataset provided us with a large amount of filtered and labelled images . It featured a comprehensive set of 600 object classes and roughly 9 million images that had been annotated with image-level annotations, but these annotated classes did not satisfy the particular needs. As a result, the image dataset with annotations to train the models were created.

1. To begin with, the system uses four classes - Refrigerator, Washing Machine, Bed and Table. An image editor tool was used\ to annotate the images with object labels. The output of this process is an xml file associated with the image file that contains the annotation details.

   **Define Classes → Image Dataset → Annotate Image Dataset → Labelling Instruction**

2. The Object Detection models utilized these annotated images in a model acceptable format as they are based on Tensor Flow. Hence another step of processing these images which converted them into. TFRecord [2] format was employed.
3. Around 500-600 images were manually annotated for each class and prepared the training dataset.

### C. TensorFlow Object Detection API

Rather than creating a model from the ground up, TensorFlow's pre-trained base models and the TensorFlow Object Detection API [2] were used.which allowed the use of transfer learning in object detection model.

Developing and training models from scratch, on the other hand, would necessitate adequate hardware.

### D. Model Building and Training

The intention was to develop an object detection model that could recognize furniture and appliances from images. The model was developed leveraging the Tensorflow Object Detection API and Tensorflow library's pre-trained SSD-Mobilenet[8] model. Using this transfer learning strategy, model could be trained using the pre-trained weights and configuration. The data was acquired via Google's open image dataset, and the complete set of images was manually labelled with an image annotation tool. The manually labelled annotation files were in.xml format, which was translated to the tensorflow-acceptable tf.record format[10]. To install all the Python files and dependencies, an Anaconda virtual environment was constructed.

The SSD-Mobilenet model configuration file, which includes the pipeline.config file and check points, was downloaded from the Tensorflow model zoo. In an 80:20 ratio, we separated our dataset into train and test sets. This train set was used to train the model, and the test set was used to evaluate it.

## IV. Results and discussion

The model performance was evaluated using Tensorboard. The training loss metrics was near to 1.5 after 5000step. The evaluation criteria is mean Average Precision. The model achieved a mAP[21] of 0.32 mAP which is a fairly good mAP score for an object detection model which has to do detection in a Realtime environment. Model was tested with different types of images and was able to identify with good confidence.

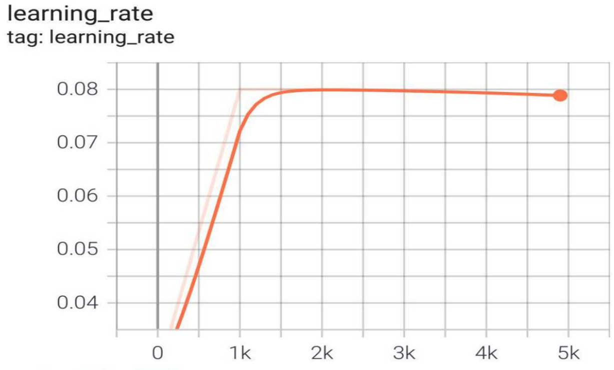

Fig. 3. Learning_rate

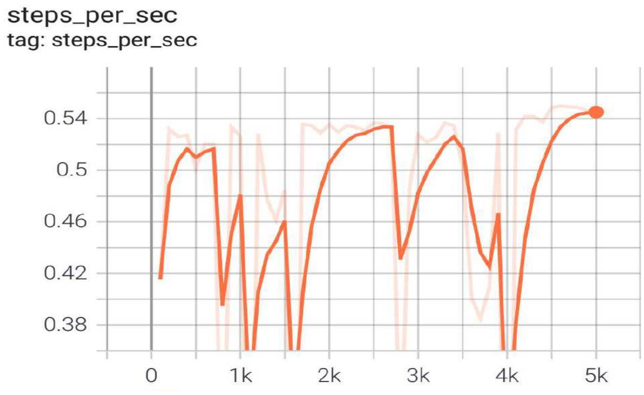

Fig. 4. Steps_per_sec

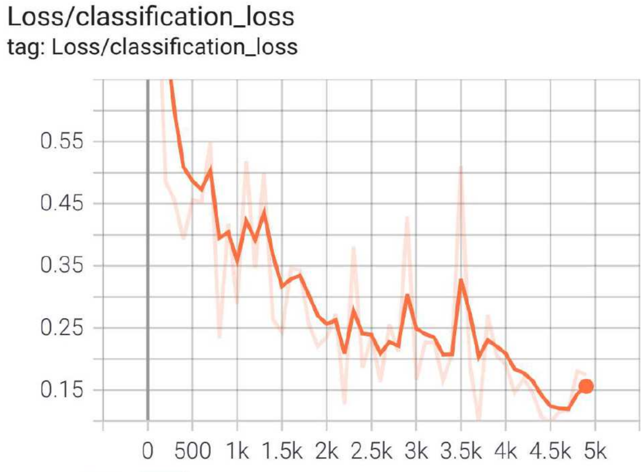

Fig. 5. Loss/classification_loss

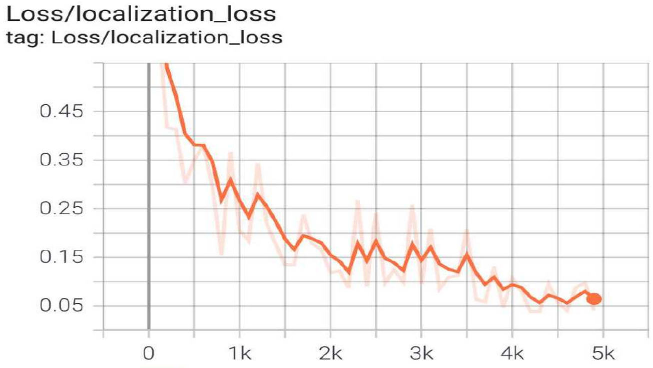

Fig. 6. Loss/localization_loss

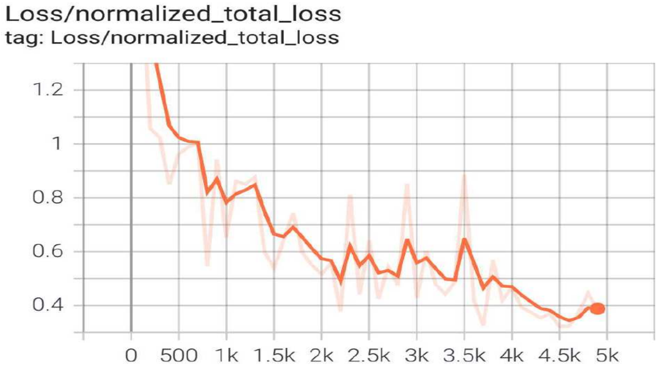

Fig. 7. Loss/normalization_total_loss

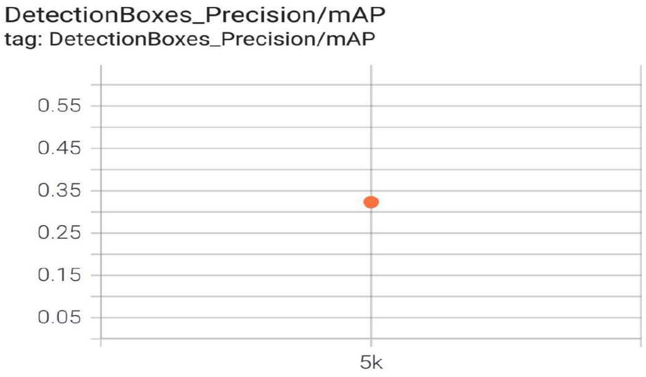

Fig. 8. DetectionBoxes_Precision/mAP

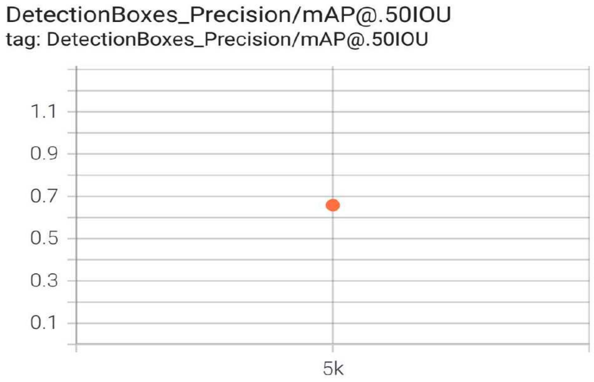

Fig. 9. DetectionBoxes_Precision/mAP

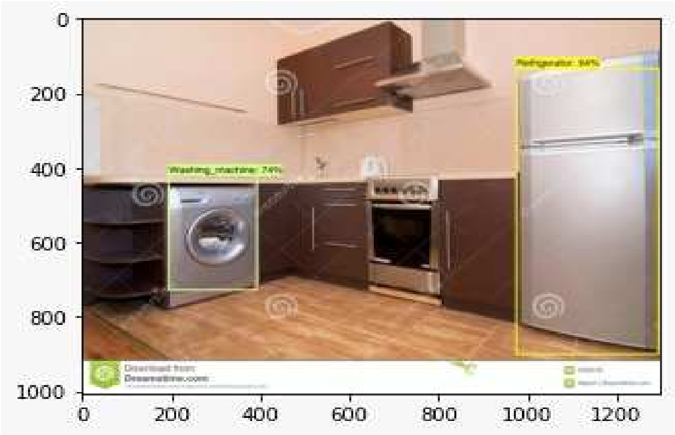

Fig. 10. Predicted object with bounding box

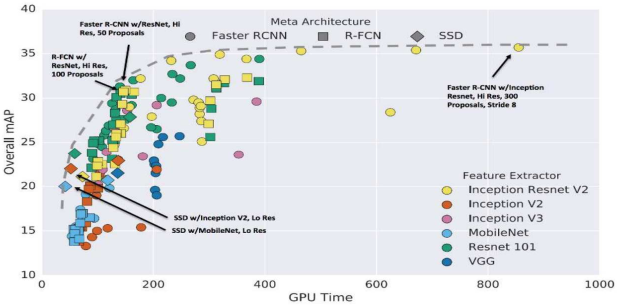

Fig. 11. Accuracy (mAP) vs interference time of different meta architecture / feature extractor combinations fos MS COCO dataset

From the above image it can be observed that SSD with MobileNet shows a mAP (mean average precision) in object detection around 20 [21]. As against the benchmarked data, our models performed better and showed an mAP of about 32 when run for 5k steps.

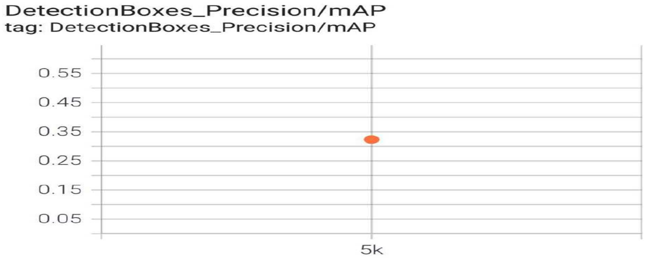

Fig. 12. Detectioboxes_Precision/mAP

- SSD with MobileNet offers the best accuracy-to-speed ratio among the fastest detectors. SSD is fast, but it performs poorly for small items when compared to others. SSD can outperform other meta-architectures with lighter extractors for large objects.
- One factor for this improved performance is the smaller number of samples per class (about 600). We had a limited number of target classes, as well as detecting items, such as washing machines, refrigerators, beds, and tables, which are larger appliances than teapots, cups, pens, and other tiny appliances.
- The amount of time spent training is considerable. Modern object recognition models are nearly entirely based on deep learning, which demands a large amount of training data. A common rule of thumb is that a few thousand image samples per class should suffice for a good model.
- Hardware restrictions - It was observed that improved hardware lowered training time.
- The amount of training the model has undergone is directly related to its accuracy.

## CONCLUSION & FUTURE SCOPE

The research and development in Advanced Computer Vision (ACV) systems can help to provide solutions to real-world problems using camera footages. The proposed system achieves real-time performance and satisfactory results, leveraging Single Shot Detector (SSD) algorithm trained with MobileNetV2 feature extraction utilizing TensorFlow for the implementation of the design. A pre-trained model trained on Microsoft-COCO dataset is employed as base model. The accuracy of the system can be improved by increasing the training data.

## REFERENCES

[1] Liu, Wei & Anguelov, Dragomir & Erhan, Dumitru & Szegedy, Christian & Reed, Scott & Fu, Cheng-Yang & Berg, Alexander. (2016). SSD: Single Shot MultiBox Detector. 9905. 21-37. 10.1007/978-3-319-46448-0_2.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Abadi, Martín & Agarwal, Ashish & Barham, Paul & Brevdo, Eugene & Chen, Zhifeng & Citro, Craig & Corrado, G.s & Davis, Andy & Dean, Jeffrey & Devin, Matthieu & Ghemawat, Sanjay & Goodfellow, Ian & Harp, Andrew & Irving, Geoffrey & Isard, Michael & Jia, Yangqing & Jozefowicz, Rafal & Kaiser, Lukasz & Kudlur, Manjunath & Zheng, Xiaoqiang. (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. Software available from tensorflow.orgK. Elissa, "Title of paper if known," unpublished.

[3] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV), 115(3):211–252, 2015.

[4] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. CoRR, abs/1311.2901, 2013.

[5] AlexKrizhevsky,IlyaSutskever,andGeoffreyEHinton.Imagenetclassificationwithdeepconvolutionalneuralnetworks.InF.Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 25, pages 1097–1105. Curran Associates, Inc., 2012.

[6] Krizhevsky, Alex & Sutskever, Ilya & Hinton, Geoffrey. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Neural Information Processing Systems. 25. 10.1145/3065386.

[7] Lin, Tsung-Yi & Maire, Michael & Belongie, Serge & Hays, James & Perona, Pietro & Ramanan, Deva & Dollár, Piotr & Zitnick, C.. (2014). Microsoft COCO: Common Objects in Context.

[8] Kumar, K. & Subramani, Goutham & Kumar, T. & Parameswaran, Latha. (2021). A Mobile-Based Framework for Detecting Objects Using SSD-MobileNet in Indoor Environment. 10.1007/978-981-15-5285-4_6.

[9] Szegedy, Christian & Reed, Scott & Erhan, Dumitru & Anguelov, Dragomir. (2014). Scalable, High-Quality Object Detection.

[10] N. Darapaneni et al., "Activity & emotion detection of recognized kids in CCTV video for day care using SlowFast & CNN," in 2021 IEEE World AI IoT Congress (AIIoT), 2021, pp. 0268–0274.

[11] Zhang, J. & Matsuzoe, H.. (2021). Entropy, cross-entropy, relative entropy: Deformation theory (a). EPL (Europhysics Letters). 134. 18001. 10.1209/0295-5075/134/18001.

[12] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556, 2014.

[13] K. Dong, C. Zhou, Y. Ruan and Y. Li, "MobileNetV2 Model for Image Classification," 2020 2nd International Conference on Information Technology and Computer Application (ITCA), 2020, pp. 476-480, doi: 10.1109/ITCA52113.2020.00106.

[14] Liao, Yuan-Hong & Kar, Amlan & Fidler, Sanja. (2021). Towards Good Practices for Efficiently Annotating Large-Scale Image Classification Datasets.

[15] N. Darapaneni, B. Krishnamurthy, and A. R. Paduri, “Convolution Neural Networks: A Comparative Study for Image Classification,” in 2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS), 2020, pp. 327–332.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[17] N. Darapaneni et al., “American sign language detection using instance-based segmentation,” in 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), 2021, pp. 1–6.

[18] Sun, Degang & Yang, Yang & Li, Min & Yang, Jian & Meng, Bo & Bai, Ruwen & Li, Linghan & Ren, Junxing. (2020). A Scale Balanced Loss for Bounding Box Regression. IEEE Access. PP. 1-1. 10.1109/ACCESS.2020.3001234.

[19] N. Darapaneni et al., “Computer vision based license plate detection for automated vehicle parking management system,” in 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2020, pp. 0800–0805.

[20] S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network," 2017 International Conference on Engineering and Technology (ICET), 2017, pp. 1-6, doi: 10.1109/ICEngTechnol.2017.8308186.

[21] J. Huang et al., "Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3296-3297, doi: 10.1109/CVPR.2017.351.

[22] K. Dong, C. Zhou, Y. Ruan and Y. Li, "MobileNetV2 Model for Image Classification," 2020 2nd International Conference on Information Technology and Computer Application (ITCA), 2020, pp. 476-480, doi: 10.1109/ITCA52113.2020.00106.

[23] Redmon, Joseph & Divvala, Santosh & Girshick, Ross & Farhadi, Ali. (2016). You Only Look Once: Unified, Real-Time Object Detection. 779-788. 10.1109/CVPR.2016.91.

[24] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid and S. Savarese, "Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 658-666, doi: 10.1109/CVPR.2019.00075.

[25] Szegedy, Christian & Liu, Wei & Jia, Yangqing & Sermanet, Pierre & Reed, Scott & Anguelov, Dragomir & Erhan, Dumitru & Vanhoucke, Vincent & Rabinovich, Andrew. (2015). Going deeper with convolutions. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 1-9. 10.1109/CVPR.2015.7298594.