IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Masked Face Detection and Voice Based Alert System for Blind People under COVID-19

Harshad N. Lokhande
Department of E&TC Engineering
Sinhgad College of Engineering
Pune, India

Sanjay R. Ganorkar
Department of E&TC Engineering
Sinhgad College of Engineering
Pune, India

***Abstract*— The COVID-19 has completely changed the world. The world is adapting newest/latest ways of living to overcome this pandemic situation. The tracing of masked and unmasked persons, can give fruitful data to identify and track the negligent people to avoid the COVID-19 infection. In this paper, new "COVID19 Asian Face Mask Dataset" (CAFMD) is created on 800 different images of masked and unmasked people from various regions of India. The module is trained with NVIDIA RTX2060 GPU on SSD Mobilenet with 88% and F-RCNN with 89.8 % accuracy for masked face detection with transfer learning. Further, the color of mask is identified and voice alert is generated to aware the blind people for precautions purpose.**

***keywords*—COVID-19, masked face, CAMFD, NVIDIA, Faster-RCNN, SSD Mobilenet, voice alert**

I. Introduction

The COVID-19 has completely changed the world. The world is adapting newest ways of living to overcome on this pandemic situation. Confirmed vaccination is not yet available to fight against the COVID-19 viruses. Hence by following certain rules and avoiding the personnel contact along with social distancing is the only solution available. The WHO insisted to wear a face mask whenever possible while moving outside. The lockdown in our country due to COVID-19 has severely affected our lives in many ways, economically, socially, emotionally and so on. To begin a life in secure way, the prevention of virus from outside world is only way. The face mask is a must to live in the community. Wearing masks is mandetory in all sectors of society namely shops, industries, schools, colleges etc. Even in some places like hospitals and crowded areas, staying unmasked is under punishment act.

The life was easy and unrestricted in the unlocking phase of COVID-19 as most of the people behaved like the pandemic ever existed. Marriages and leisure activities were almost back in full-swing with turning shopping malls and streets as most dangerous hotspots [1].

The tracing of masked and unmasked persons, can give fruitful data to identify and track the negligent people to avoid the COVID-19 infection. In India, people are wearing a variety of masks made by materials like cotton, silk fabric, surgical papers and also medically proven N-95 Masks [2]. Even few people were used to put a hand or handkerchief on their faces. Considering all above situations, the masked person detection is critical task in India. There are smart systems available using deep learning to detect the masked faces using labelled datasets [3][4][5], but these are not well versatile with Indian mask types and wearing patterns. Hence, we created own face mask Dataset "COVID19 Asian Face mask Dataset" (CAFMD) with 1500 different images of masked and unmasked people from various regions of India as well as some Asian countries. The proposed method achieved masked and unmasked persons detection ratio with 80.03% accuracy with several types of masks.

II. Review Work

The object can be detected using traditional methods based on edge and grey value of face from an image [6]. The edges will define the face pattern with prior information about face model. The face detection was improved with Advanced Viola Jones Method [7], with false detection under factors like illumination, brightness change and face shape orientations. The model didn't work under dark light conditions. Based on the feature, the mask can be detected with regular computer vision methods like Haar cascade, HOG [8] and machine Learning algorithms with Linear SVM.

These challenges can be overcome by deep learning methods and computer vision algorithms [9]. The models are trained with supervised labelled data [10]. The object detection irrespective of illumination, shape and rotation can be achieved with CNN [14]. The training of model from the scratch is very complicated task. The use of transfer learning on pre-trained models saves the time and provides almost equivalent accuracy in detection of custom objects. The Retina Facemask was proposed by Jiang et al. [11] proved high accuracy in mask detection with ResNet and MobileNet as base models.

According to [9] the detection of face is prior task for knowing the mask. The convolutional Neural Network is backbone of deep learning for object detection. Instead of objects manual feature extraction, CNN with sliding window gives exact number of features essential for object detection.[11] Ejaz et al. demonstrated non-masked face detection using approach of Principal Component Analysis, with accuracy up to 99% with slow frame rate [12]. The frame rate is increased with Real-Time Face Mask Detection using YOLO V3 network and Region Proposal Network [13][14].

Though CNN architecture gives good results, still it is slow to process [15] with sliding a window at every frame for convolution. There is no exact control on aspect ratio of image since CNN requires the fixed size of image. So, if size of image is increased, more sliding window with more pyramid layers can lead to an increase in the errors. Separate feature extraction leads to a tedious job which can be made more flexible by implementing deep learning for exact classification between different objects in a scene.

III. Methodology

The masked person detection systems are deployed using pretrained VGG16 based SSD and F-RCNN architectures. The Faster-RCNN is computed with COVID19 Asian Face Mask Dataset: CAFMD a priori. The detection of masked person focuses on bounding box method from Faster RCNN. As shown in figure 1, the bounding boxes with (x, y) as locating coordinates defines the class of an object from that image. To predict the correct object, ground truth of bounding box of labelled image has to be matched with the predicted bounding box. While evaluating the mask detection performance, the score of Intersection over Union (IoU) of detected mask should be more than 0.5 to accept as positive detection.

In Single Shot Detection Method (SS), the image is applied on pretrained VGG base network by replacing POOL and CONV layers with new CONV layer of size 1x1, 3x3 kernels to reduce dimensionality with high content features. With multibox algorithm, each cell from feature map are mapped for ground truth bounding boxes. The SSD considers two types of loss function during learning process, the confidence loss and location loss. The cross-entropy loss is helpful for measuring the confidence of class of the object predicted under bounding box. The SGD optimizer preferred for an end-to-end training. For last layers the Adam optimizer gives more accurate detecting in fine tuning of networks.

The feature extraction from image can be done by CNN by removing final layer (FC). Then with restructuring the CNN with additional pretrained weights from SSD or Faster R-CNN, the model gets transferred to learn new objects.

A. *Dataset*

The existing masked people detection datasets shows good results on sharp images with surgical masks, but fails to detect faces with other type of masks. Hence new dataset is developed as “COVID19 Asian Face Mask Dataset” (CAFMD) with mixed type of images on variety of masks under various situations like single person or crowded area, both masked and unmasked people together [16]. In early stage of lockdown, the tracking of migrant worker people was one of the major issue faced by government [17] [18]. So, in our dataset we considered this as a challenging task, to collect such images and use for mask detection. Using the technique of image web scraping [15], the images were collected from sources like PTI (Press Trust of India) [16], The Times of India Images [19] and the Google. The major challenge was to identify images of masked people due to COVID19 threat. Once images are scrapped, all are separately analyzed and sorted according to criteria as shown in table 1:

TABLE 1: DISTRIBUTION OF MASKED IMAGES

Type of Image	% of images in dataset
Masks with fabric	40 %
Surgical Masks	10 %
Unmasked people	40 %
Both masked and unmasked	10 %

B. *SSD*

For mask - detection, two different methods are implemented. The single' shot detector (SSD) with two stage mechanism [20], overcomes the disadvantage of CNN’s slow execution due to sliding windows effect for convolution. In this project the backbone of SSD relies on pretrained network VGG 16 for feature extraction. The input image is parsed to VGG 16 [21]. The fully connect layer is removed by preserving extracted semantic features. The SSD head consists of one or more convolutional layers mapped with this pretrained model. The output is interpreted as the bounding box and the name of objects class with exact location in the input image. The architecture of SSD is shown in figure [1].

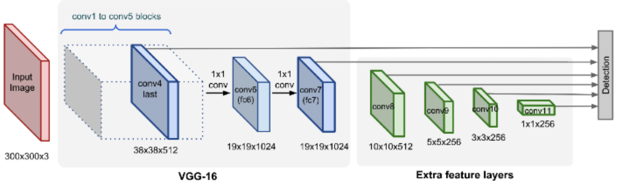


Figure 1. SSD Architecture [21].

The SSD splits the input image in grids and each grid corresponds to detect and locate object in that region. If object is present, then it is considered with spatial position and predicted class of object, else it goes as background class and ignored. The detection mechanism is shown in figure [2], by the 4×4 grid splitting an image which provides output as location with name of object it detects.

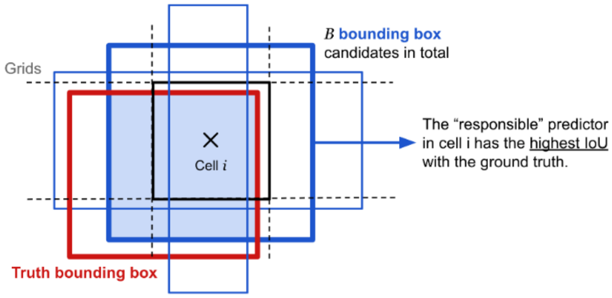

B bounding box candidates in total
Grids
Cell *i*
The “responsible” predictor in cell i has the highest IoU with the ground truth.
Truth bounding box

Figure 2. Grid concept and Anchor box for masked person detection [21].

The anchor box [2], shown in fig. 2, simplifies the issue of locating and identifying multiple objects in same grid. The SSD works with matching of Anchor box with bounding box for each ground truth from the input image. While inferring the output, probably not all masked persons from same category are with same size in bounding box, so, SSD uses pre-defined aspect ratios of anchor boxes and variations in zoom level. In transfer learning and feature extraction, the VGG16 architecture is used by taking outputs before the final POOL layer. The output is 512×7 X7 which when flattened then results to feature vector of size 25,088. So, while using VGG16 for feature extraction, the N has to set with the size of total number of images in training dataset.

C. *Faster RCNN*

The faster RCNN works on TensorFlow as backbone. The input dataset images are converted in TensorFlow’s encoding functions to know each attributes of object from an image. Each converted row from RCNN contains, a class label, x and y coordinates with starting of bounding box, along with x and y coordinates at the end of bounding box. The Faster R-CNN contains the Region Proposal Network (RPN) addition with Fast R-CNN as the detector network [22]. The input image is passed over the Convolutional Neural Networks (CNN) to extract the features. Then, the network passes the output of the ROI pooling layer through two Fully Connected (FC) layers to provide the input of a pair of FC layers that one of them determines the class of each object and the other one performs a regression to improve the proposed boundary boxes

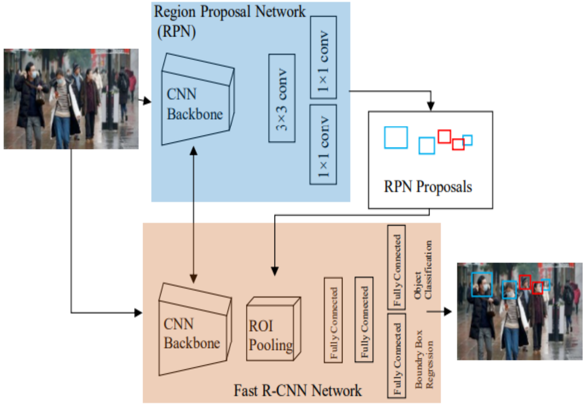


Figure 3. A schematic architecture of the Faster R-CNN

D. *Identifying Color of Detected Mask*

The color detection of mask was a challenging task. The detected objects label is compared with global ID. If the confidence level of detected object is greater than 0.65, then only the bounded region of the detected object is cropped. To identify the dominant color, the K-Means clusters are used with getting the centroid of most dominant colors from the cropped region. The result is compared with RGB color_pallete and identified decimal value is returned. The decimal value is converted to rgb_triplet in WebColours format. The compared string is returned as color of detected mask

E. *Voice based alert mechanicsm*

The masked person is detected, the color of mask is retrieved and then it is passed to sound engine. The extracted color with mask detection extracted in string format. The Text to Speech engine (TTS), convert this string into voice prompted alert for blind people. The color will be prompted in real-time. So, this system provides a real-time sound alert of detected mask and the identified color.

IV. Experiment Results

The masked person detection deployed using pretrained VGG16 based SSD and F-RCNN architectures. On prior, the Faster-RCNN is computed with

A. *Experiment Setup*

The training depends on Stochastic Gradient Decent (SGD) with learning rate $\alpha = 10^{-3}$, momentum $\beta = 0.8$, and epochs =250, as optimization algorithm. The complete training was done on NVIDIA GeForce RTX 2060. The dataset has been split into a train, test set with 942 masked and 685 as unmasked labels on 618 images. The pretrained weights are used with transfer learning and feature extraction, the VGG16 as base architecture is used by taking outputs before the final POOLING layer. The output is 512 x 7 x 7 which when flattened then results with feature vector of size 25,088. Hence, using VGG16 for feature extraction, the N has to set with the size of total number of images in training dataset.

B. *Training*

The complete project was trained on two separate algorithms. Following figure 4 is showing training loss and localization errors achieved during the training process.

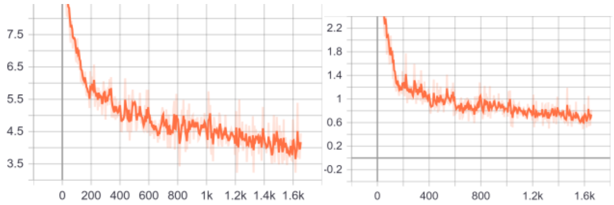


Figure 4. Training Loss and Localization loss values received on Faster RCNN Model.

C. *Testing*

Once the model is trained with 0.2 to 0.3 % loss in accuracy, the frozen model is used for testing images and real-time camera footage. Following are few results obtained from various situations, angles and scenes. The following figure 5 is showing the multiple masked faces detected using MobileNet SSD. Here from three masked faces, it detected two face correctly.

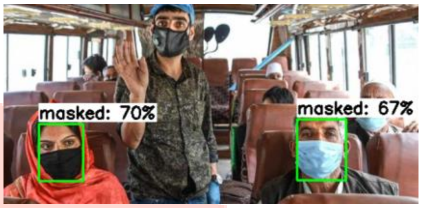
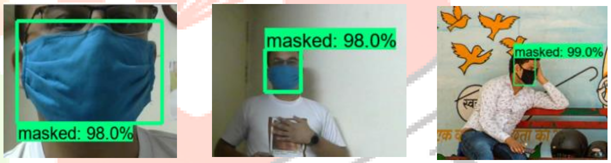


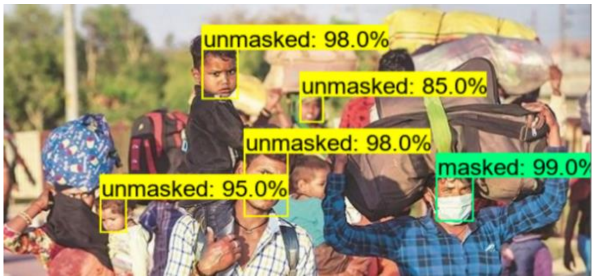
Figure 5. Multiple face detection using MobileNet SSD



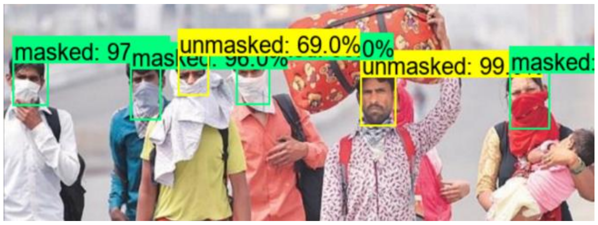
(a) (b) (c)

Figure 6. Results with Faster RCNN with web camera

Further study, includes the detection of masked face with Faster-RCNN network. The accuracy of results achieved with Faster RCNN is more than MobileNet. The figure 6, shows three different scenarios. a) partial face but with mask, b) detected masked face even from long distance with FRCNN method, c) the person posing in cross sectional view, detected as masked with true case.



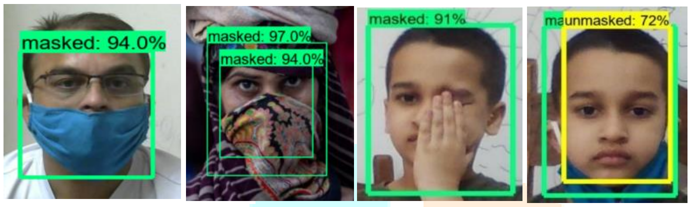
(a)



(b)

Figure 7. Multiple object detection from single image

In the figure 7(a), the prediction accuracy of correct class is almost 99.0 %. In 7(b), the 3rd person from left side, having mask partially open at nose, and our model correctly identifies him as "unmasked". So, the prediction accuracy is high in Faster -RCNN than Mobilenet SSD architecture.



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

Figure 8. False predictions with Mobilenet SSD

Though masked-unmasked detection ratio is good, there are few observations in which our model fails to correctly identify the masked face. The figure 8, is showing few false predictions in various scenarios.

From left, figure a., in which the partially face is unmasked, but it detected as masked with 94% confidence. The figure b, is having challenge of overlapping anchors for same face masked person. The figure c, detected child as masked with hand kept on his face, with the face covered by hands partially. And in figure d, showing overlapping in class due to IoU @ 0.5 for both the classes.

D. *Evaluation Metrics*

The precision and recall metrics are implemented for evaluation of results.

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (1)$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

The equation 1 and 2 gives precision and recall values evaluated on test images. The following figure 9 shows the confusion matrix for getting accuracy of results for detecting the masked face.

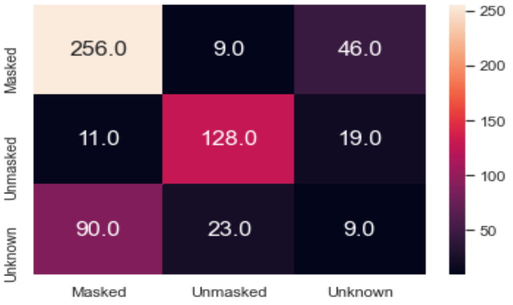


Figure 9. Confusion Matrix to evaluate the results.

The evaluation was done on 130 test images with Faster RCNN method. The Y- axis is actual label and X- axis is an predicted labeled image.

Discussions

For object detection Mean Average Precision (mAP) considered as benchmark. The mAP calculates Intersection of Union (IoU) for each overlapping bounding box. Typically, mAP is considered to mAP@0.5, means in order to detect object with labelled output it should have 0.5 IoU with ground truth.

As SSD uses gradually decreasing feature map system for resolutions, it fails to identify masked persons from far away distance. This issue could be resolved by increasing size of input image.

Faster-RCNN uses a "fix aspect ratio", which mean, the size of each input image will remain same in dissimilar mode even after pre-processing. This practically makes batching during training as impossible case, since a batch tensor has a fix shape of [num_batch, height, width, channels]. Hence its mandatory to keep batch size =1 for Faster RCNN training.

Whereas the SSD model, uses a "normal" resizer with regardless of the input image So in result all pre-processed images will be converted to the same size, which allows them to be use in batches.

The table II shows improved results for detecting masked faces with 0.82 as recall @ 0.5 IoU. While, still there is scope for improvement in precision as 0.717 @ 0.5 IoU. By increasing the training images and labeled dataset, the precision value can be improved.

With results obtained in confusion matrix, 90 objects were masked, but detected as unknown, as the images are with low resolution and small bounding box are not able to detect with multiple objects. This limits the scope of SSD and Faster RCNN algorithm. For detecting multiple objects in small size, may need to go with YOLO-5 method for object detection.

Conclusion

In COVID-19 situation, having a mask on face was the only solution to avoid the infection. So, this system helps to detect the unmasked person with 91.9 % accuracy in different scenarios. The class detection compared using two methods Mobilenet_SS_V2 and Faster RCNN based on VGG16 architecture. The results with Faster RCNN are more accurate and robust even for farthest objects also. Whereas the Mobilenet_SSD is faster to train and easy to deploy on lower end systems with low detection accuracy up to 86.1 %. Further unmasked person detection will raise an audio alert to know his presence within scene. This will help to keep a track of unmasked persons.

TABLE II : Precision – Recall on F-RCNN

Category	PRECISION_@0.5IOU	RECALL_@0.5IOU
Masked	0.717087	0.823151
Unmasked	0.800000	0.810127

REFERENCES

[1] (www.dw.com), D., 2021. Coronavirus in India: Is public negligence causing surge in cases? | DW | 31.07.2020. [online] DW.COM. Available at: <https://www.dw.com/en/coronavirus-in-india-is-public-negligence-causing-surge-in-cases/a-54394810> [Accessed 31 March 2021].

[2] MacIntyre CR, Chughtai AA. A rapid systematic review of the efficacy of face masks and respirators against coronaviruses and other respiratory transmissible viruses for the community, healthcare workers and sick patients. Int J Nurs Stud. 2020 Aug;108:103629. doi: 10.1016/j.ijnurstu.2020.103629. Epub 2020 Apr 30. PMID: 32512240; PMCID: PMC7191274.

[3] Howard, Jeremy et al., (2020). Face Masks Against COVID-19: An Evidence Review. 10.20944/preprints202004.0203.v1.

[4] Matthias, Daniel & Managwu, Chidozie. (2021). FACE MASK DETECTION PAPER (1). 10.13140/RG.2.2.18493.59368.

[5] Loey, Mohamed, et al. “Fighting against COVID-19: A Novel Deep Learning Model Based on YOLO-v2 with ResNet-50 for Medical Face Mask Detection.” Sustainable Cities and Society, vol. 65, 2021, p. 102600., doi:10.1016/j.scs.2020.102600.

[6] Wang, Zhongyuan, and Guangcheng Wang. “Masked Face Recognition Dataset and Application.” Computer Vision and Pattern Recognition, 20 Mar. 2020.

[7] Huang, J., Shang, Y. & Chen, H. Improved Viola-Jones face detection algorithm based on HoloLens. J Image Video Proc. 2019, 41 (2019). https://doi.org/10.1186/s13640-019-0435-6

[8] LI, CHUNXIANG, et al. “A Novel Algorithm for Fast Human Detection Based on HOG.” DEStech Transactions on Computer Science and Engineering, no. aiea, 2017, doi:10.12783/dtcse/aiea2017/15021.

[9] Kibria, Sakib B., and Mohammad S. Hasan. “An Analysis of Feature Extraction and Classification Algorithms for Dangerous Object Detection.” 2017 2nd International Conference on Electrical &amp; Electronic Engineering (ICEEE), 2017, doi:10.1109/ceee.2017.8412846.

[10] Shaoqing R, Kaiming H, Girshick R, Jian S (2015) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39:1137–1149.

[11] Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017

[12] Jason B A Gentle Introduction to Transfer Learning for Deep Learning. https://machinelearningmastery.com/transfer-learning-for-deep-learning/

[13] Singh, S., Ahuja, U., Kumar, M. et al. Face mask detection using YOLOv3 and faster R-CNN models: COVID-19 environment. Multimed Tools Appl (2021). https://doi.org/10.1007/s11042-021-10711-8

[14] Tsai, An-Chao, et al. “Efficient and Effective Multi-Person and Multi-Angle Face Recognition Based on Deep CNN Architecture.” 2018 International Conference on Orange Technologies (ICOT), 2018, doi:10.1109/icot.2018.8705876.

[15] Hosseini, Hossein, et al. “On the Limitation of Convolutional Neural Networks in Recognizing Negative Images.” 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), 2017, doi:10.1109/icmla.2017.0-136.

[16] Photo Gallery, Press Trust of India, ptinews.com/ptigallery/photonewuser.aspx?pcat=3.

[17] Khanna, A. (2020). Impact of Migration of Labour Force due to Global COVID-19 Pandemic with Reference to India. Journal of Health Management, 22(2), 181–191. https://doi.org/10.1177/0972063420935542

[18] Irudaya Rajan, S., Sivakumar, P. & Srinivasan, A. The COVID-19 Pandemic and Internal Labour Migration in India: A ‘Crisis of Mobility’. Ind. J. Labour Econ. 63, 1021–1039 (2020). https://doi.org/10.1007/s41027-020-00293-8

[19] “Web Scraping: Latest News, Videos and Photos of Web Scraping: Times of India.” The Times of India, Search, timesofindia.indiatimes.com/topic/web-scraping.

[20] Loey, Mohamed, et al. “Fighting against COVID-19: A Novel Deep Learning Model Based on YOLO-v2 with ResNet-50 for Medical Face Mask Detection.” Sustainable Cities and Society, vol. 65, 2021, p. 102600., doi:10.1016/j.scs.2020.102600.

[21] Weng, Lilian. “Object Detection Part 4: Fast Detection Models.” Lil'Log, 27 Dec. 2018, lilianweng.github.io/lil-log/2018/12/27/object-detection-part-4.html.

[22] Ren, Shaoqing, et al. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, 2017, pp. 1137–1149., doi:10.1109/tpami.2016.2577031.