Safety Helmet Detection using Deep Learning

Harsh Shah, Lalithambal Swaminathan
Wilfrid Laurier University
Waterloo, ON, Canada
{shah5610, swam5140}@mylaurier.ca

***Abstract*—Object detection, a major challenge in Computer vision generally deals with the detection of target objects of different classes in a graphic view of an input image. Image understanding, object tracking and video analysis have a close relationship with object detection. This fundamental technique has a lot of applications in various industries and use cases ranging from productivity to personal security. In this paper, we discuss Faster R-CNN with Inception V2 as our model that we trained and fine tuned on Hard Hat workers dataset which is publicly available. As, the published dataset is relatively new, no prior research has been done on it with the model that we have used. So, we conducted various experiments by tuning the configuration parameters and evaluating with different metrics. After final training, we achieved 64% of mean average precision value with detection results having high confidence threshold.**

***Index Terms*—Construction Workers, Deep learning, Object detection**

I. Introduction

Computer vision is a field of artificial intelligence that teaches computers to read and understand the visual world. Using digital images, videos from cameras, and deep learning models, machines can accurately recognize and distinguish objects and then react to what they see. Object detection has been an active area of research for several decades. A wide range of applications, notable technological breakthroughs are a couple of the key factors for the significant advancement of object detectors in recent years.

The domain of computer vision embodies several recognition problems, like image classification, object detection, semantic segmentation, and instance segmentation. Image classification is a particular technique responsible to categorize semantic classes of different objects in an image. On top of that, the task of object localization is to localize that major object of interest with a bounding box, While object detection identifies multiple object categories along with the accurate locations of all those objects and shows detected objects via bounding boxes. This fundamental computer vision technique acts as a backbone for image understanding and to perform high-level computer vision tasks such as event detection, object tracking, object segmentation, scene text detection, image captioning, etc. An object detection algorithm can be considered good if it can attain a robust semantic understanding with spatial data about input image.

Predominantly, object detection has been employed in numerous computer vision applications like face detection, personal safety, security surveillance, pedestrian detection, license plate and traffic signs detection, and video analysis. However, concurrently due to substantial variations in lighting conditions and occlusions, different poses and, varied viewpoints, object detection with accurate localization has become a difficult problem and, it is quite hard to accomplish it with outright perfection. So, these challenges have contributed significantly to give the rise in attention to this field in recent years.

Its been more than seven decades since a mathematical model in the 1940s, McCulloch Pitts Neuron showing biological neuron was published that established the foundation of neural network and deep learning. These neurons are connected in layers that build the neural network with complex and deep structures and are called deep models. In the second half of the last decade, deep models were used in pattern and speech recognition by researchers that showed significant performance improvement and, these breakthroughs allowed them to get deployed in more commercial industries. After 2013, most of the computer vision started using neural networks, and later in 2016, natural language processing also switched to it. Large-scale training data with annotations, Rapid development of a high-performance system for parallel computing like GPU clusters allowed neural networks to show their high learning capacity with significant reduce in training time. Emergence of efficient training strategies such as data augmentation, batch normalization, dropout, and study of the varied design of neural network structures such as GoogLeNet, AlexNet, Residual Net (ResNet), and Visual Geometry Group (VGG) extensively helped to enhance the overall performance.

Convolutional Neural Network is a sub-category of neural networks and therefore has all the fundamental properties of neural networks. Its architecture is hierarchical and composed of different layers. Generally, CNN embodies several feature maps, and in that each pixel functions as a neuron. Typically, there are four types of layers named as a convolutional layer, the pooling layer, non-linear activation function (e.g. ReLU), and fully connected layer, and data is represented through these layers. These different blocks of operations (convolution, pooling, and activation function) are illustrated in Fig. 1.

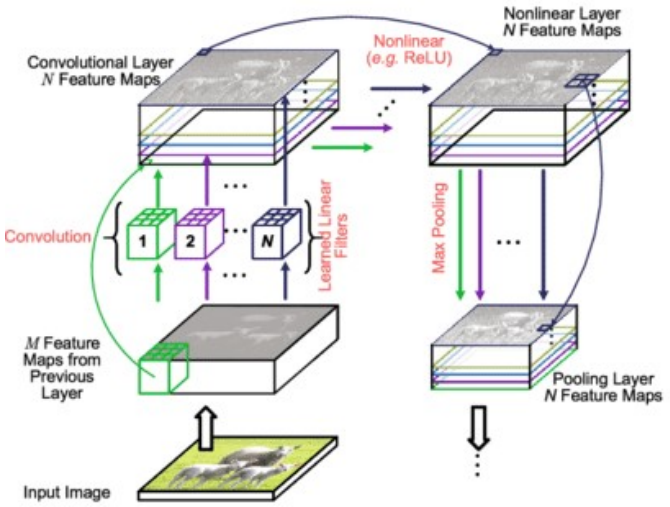

Fig. 1. Convolutional Neural Network Architecture[1]

II. Related work

A. Construction safety helmet detection

Currently, the previous studies related to safety helmet detection can be split into three parts, with the first one being the Sensor-based object detection, the second being the Machine learning-based object detection and, the third being the Deep learning-based object detection. Sensor-based object detection usually uses the Radio Frequency Identification (RFID)tags [2][3] and locates the helmets worn by the workers to ensure their safety in real-time by monitoring them with the help of a portal[4]. However, this method of detecting the Helmet has the problem of ensuring the safeness of workers because RFID readers have limited bandwidth and, the workers who are beyond the RFID range cannot be monitored for their safety.

The machine learning-based object detection is usually automatic to obtain the attributes of the helmets and detect them[5][6]. One such method performs the detection in two steps which utilizes the Circle Hough Transform with the combination of the histogram of oriented gradient from the image. The end result obtained is very accurate yet this type of detection can detect safety helmets with defined colors and it fails to distinguish the workers wearing the same sized hats with the same defined color and when the workers did not show up their faces towards the camera in the construction site. So this type of traditional detection has very good accuracy despite having a complex design but suffers from generalization ability.

B. Object Detection using deep learning

With advancements in Deep learning-based object detection, the methods for object detection are performed by two types, one by using the one-stage detection method and the two-stage method with region suggestion.

The two-stage detection algorithm region-based and typically begins with the extraction of object candidate boxes(proposals) as a set by the selective search. One such is the R-CNN detector which extracts the features and, a fixed-size image is rescaled and given as input into the CNN model. Then, SVM classifiers predict the object's presence and categorize them accordingly.

Fast Regions with CNN feature (Fast R-CNN) and Faster-R-CNN are classified under the two-stage detection algorithms, with Faster-R-CNN having the best detection performance so far.

One stage object detectors does not maintain a separate branch for generating proposals and ponder over all the locations of an input image as prospective objects and further attempt to categorize each region of interest (ROI) into either background or a predefined object class. SSD, Retinanet, YOLO, YOLOv2, and YOLOv3 come under this One-stage detection where YOLOv3 is the most effective amongst them. By utilizing YOLOv3 and DenseNet in model parameters - YOLO-Densebackbone convolutional neural network, the helmet detection accuracy is increased by 2.44% than the traditional YOLOv3. This improved model also deals with the detection of helmets which are partially stained and input image of low resolutions[6].

III. Algorithm

The region proposal-based framework focuses on the region of interest by first scanning the whole scenario. Among the existing model available, the Overfeat model predicts the bounding box from the feature map (locations) by inserting CNN into the sliding window.

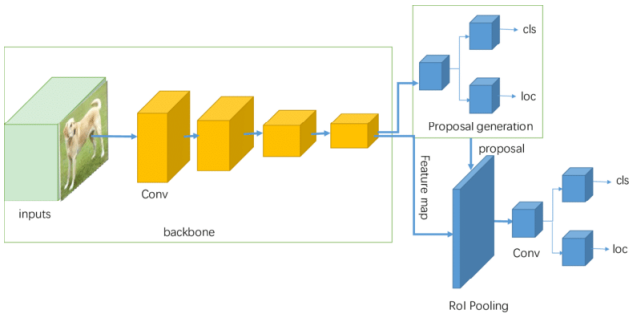

Fig. 2. Region Proposal based Framework Architecture[7]

A. Faster R-CNN

To address the issues caused by conventional techniques like selective search and edge boxes that use visual signals and are difficult for detectors to learn the data-driven approach, Faster R-CNN was proposed shortly after the release of Fast R-CNN. A fully convolutional layer called Region Proposal Network (RPN) which creates proposal sets on each feature map location by working on random images and is used in Faster R-CNN. The derived feature map yields feature vectors that are fed into the classification layer and then into a bounding box regression layer for localization leading to Object detection.

This model consists of two modules where the first one is a fully convolutional network that proposes regions, and the second is the Faster R-CNN detector that uses the proposed regions, and these two modules form the unified network for Object detection. The region proposal network (RPN) is for generating region proposals and a network using these proposals to detect objects.

RPN takes an input image of any size and provides a set of rectangular object proposals as output where each proposal has an object-ness score. A small network that takes as the input of an n × n window of the feature map slid over the output convolutional feature map. This feature is given as an input into two fully connected layers out of which one is a box classification layer, and another is a box regression layer. This typical architecture is implemented with n x n convolutional layer followed by 1x1 convolutional layers.

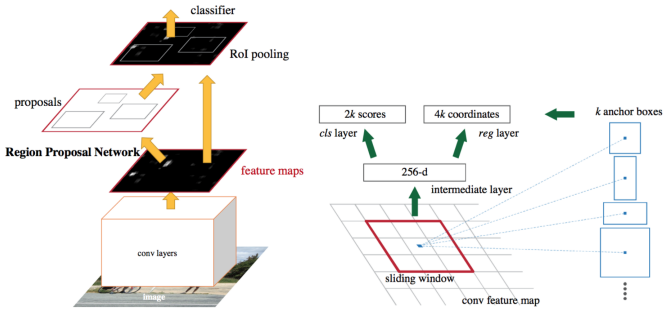

Fig. 3. Left side: Faster R-CNN Architecture, Right side: Region Proposal Network[8]

Anchors: Multiple region proposals are simultaneously predicted at each sliding-window location, and k denotes the maximum number of possible proposals for each location.

ROI Pooling: As the Regional Proposal Network produces proposed regions of different sized convolutional feature maps, Region of Interest Pooling handles the problem to work with features of different sizes by reducing the feature maps into equal sizes. ROI pooling splits the input feature map into a decided number of roughly equal regions, followed by Max pooling on each region.

B. Inception V2

Inception network was complex, developed to increase the performance when compared to other CNN classifiers. Inception v2 was developed to reduce 'representational bottleneck' which occurs when the dimension of the input image is reduced drastically causing loss of information.

In this version, a 5x5 convolution is factorized to two 3x3 convolution layers to increase the computational speed and the former is 2.78 times expensive than the latter convolution operation. And also combining 1 x n and n x 1 convolutions by factorizing convolutions of filter size n x n is found to be 33% cheaper than a 3x3 convolution.

In order to remove the representational bottleneck, filter banks present in the module were made wider instead of having it deeper. Having the deeper filter would result in an extreme reduction in dimension leading to loss of information.

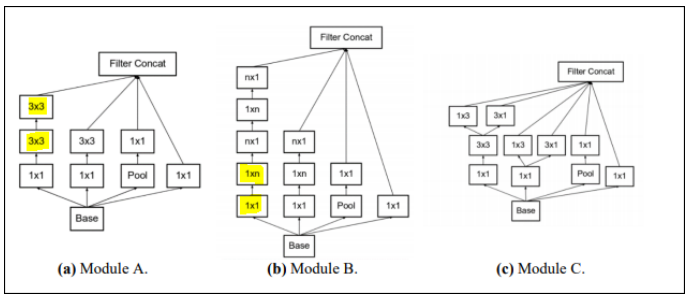

Fig. 4. Module Inception V2 [9]

IV. Dataset and Evaluation Metric

A. Dataset

Our dataset for this project is publicly available on [10]. It is hosted by ROBOFLOW named as Hard Hat Workers. As per the documentation, "the Hard Hat dataset is an object detection dataset of workers in workplace settings that require a hard hat. Annotations also include examples of just "person" and "head," for when an individual may be present without a hard hat." So, this dataset has in total 3 object classes.

This dataset is shared by Northeastern University in China in the month of April,2020. The total number of images including train and test are 7041. This dataset is available in various formats as per user's requirement. For our work, as we worked using TensorFlow API, we used the tfrecord file format for training.

As per our observations, the images in the dataset has huge differences in terms of helmet viewpoints. For any of the image, isCrowd variable value is always false. The image data ranges from small to large pixel coverage of helmets within images which is the actual scenario in real world application.

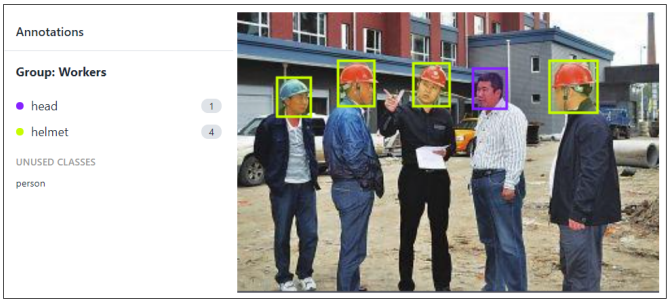

Fig. 5. Hard Hat workers dataset [10]

B. Evaluation Metrics

In general, there exist numerous evaluation metrics that are used for different algorithms to measure performance. For evaluating the performance of object detectors, there are mainly three criteria[11]:

i. Precision

TABLE I
SAMPLE IMAGES WITH ANNOTATIONS INFORMATION

Image ID	Width	Height	Class	Xmin	Ymin	Xmax	Ymax
1654	416	416	Helmet	296	9	37	37
1658	416	416	Helmet	399	8	15	38
1777	416	416	Helmet	226	124	41	54
4395	416	416	Head	389	204	16	18
4399	416	416	Head	158	208	9	15

Fig. 6. Hard Hat workers dataset

ii. Recall
iii. Detection speed in Frames Per Second(FPS)

As in our work, only image dataset is used and no video data is there, Average Precision (AP) is preferred for Object detection which is derived from precision and recall and originally introduced in VOC2007. Intersection over Union (IoU) is used to measure the object localization accuracy.

IoU threshold is predefined with a value, say 0.5. If IoU between the predicted box and the ground truth box is greater than the set threshold, the object will be identified as "successfully detected", otherwise identified as "miss". As per our research, mAP@0.5IoU has been a standard metric for detection problems so we will be using this mainly to evaluate our performance.

V. EXPERIMENTAL EVALUATION

For our experiments we have splitted the dataset with 70% images for training which is in total 5269 images, 13% for test and 12% for the validation set. We have used 'transfer learning' concept in which we needed to customize the configurations and use the pre-trained weights to fine tune the model for our helmet detection task.

To customize the configurations of the parameters, we did some research and analysis and understood few parameters and their importance. These details are summarized in Table II with the final values of those parameters that we selected to fine tune our model.

TABLE II
HYPER-PARAMETERS DEFINED FOR OUR FINAL MODEL

Hyper-parameters	Tuned value
batch_size	1 (Due to restricted GPU usage)
initial_learning_rate	0.002
num_steps	120000
metrics_set	coco_detection_metrics
first_stage_nms_iou_threshold	0.7
score_converter	SOFTMAX
second_stage_post_processing / / iou_threshold	0.6
use_dropout	false

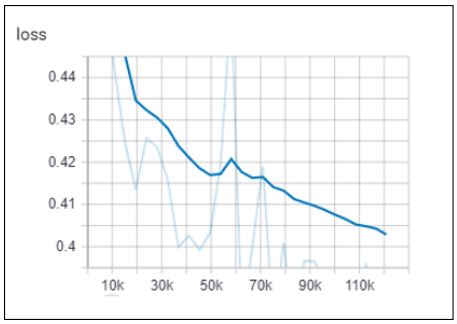

Fig. 7. Validation Loss

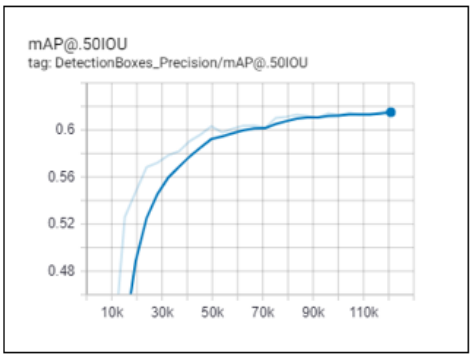

Fig. 8. Metric - Mean Average Precision

In Fig. 7., it can be seen that the validation loss after training steps of 120000 was 0.40 and with that we got 64% of mAP for IoU >= 0.5 for detection boxes as can be seen in Fig. 8.

In Fig. 9, average recall values are shown for different object sizes and as per our observation, these values are somewhat similar which means model has generalized the learning and so different sizes of helmet (pixel wise coverage) in the image

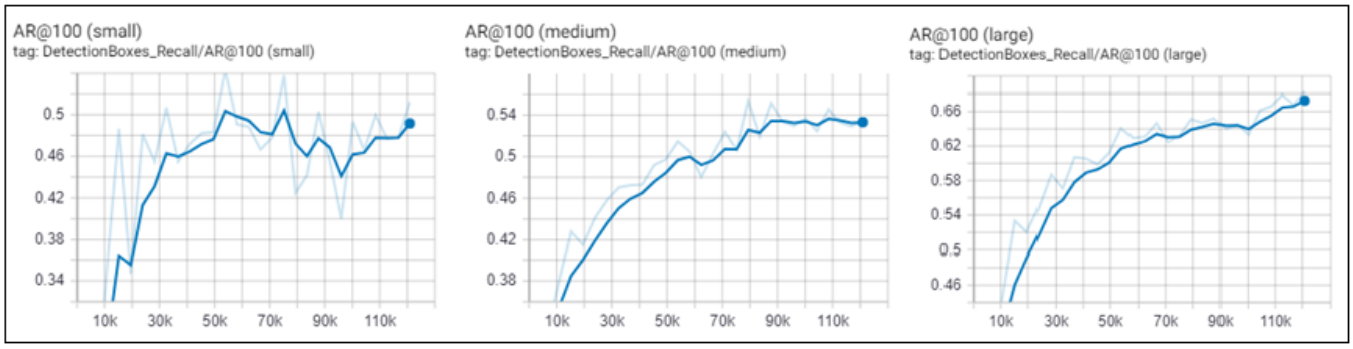

Fig. 9. Metric - Average Recall

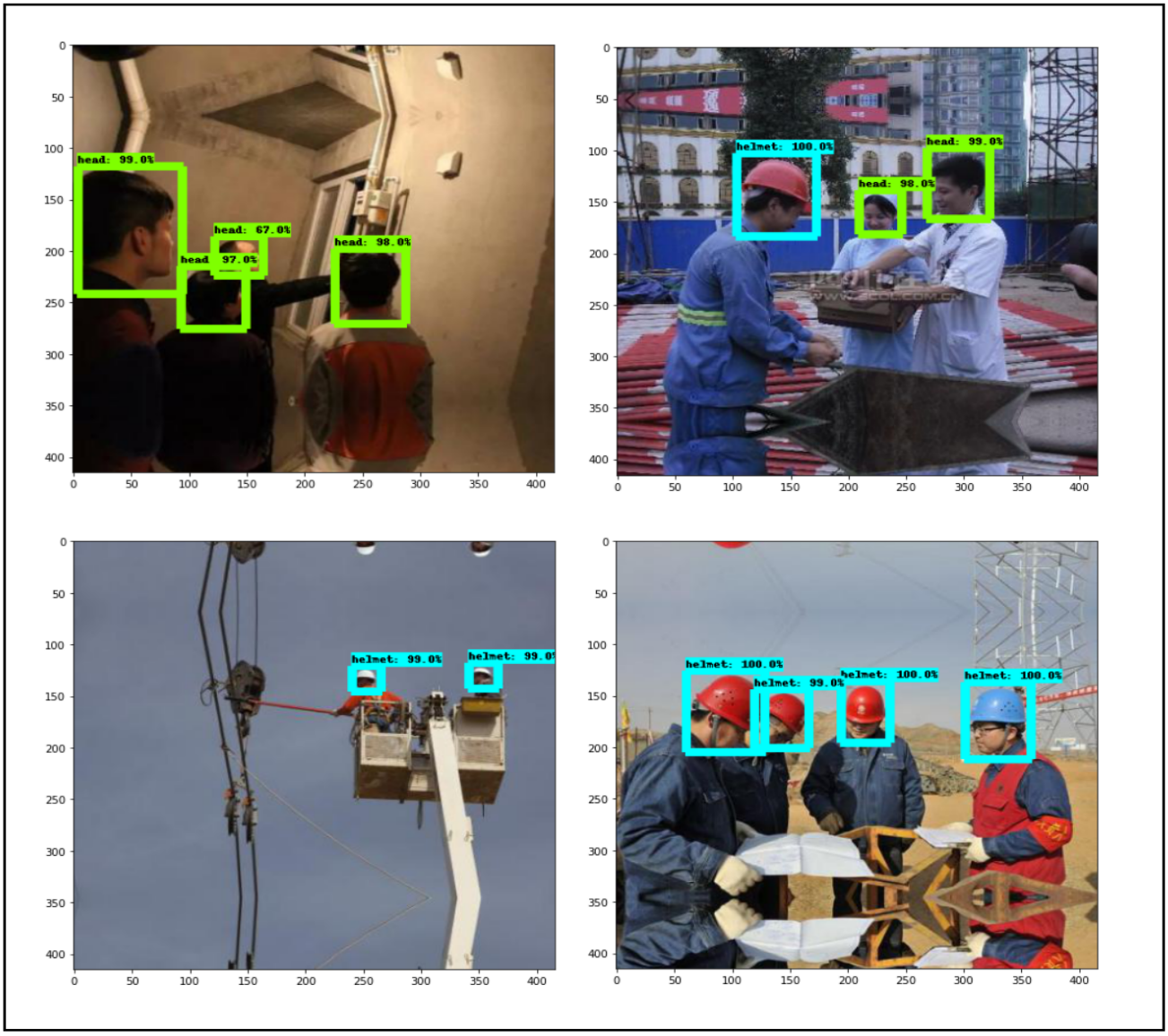

Fig. 10. Detection Results

should be detected with similar confidence threshold and these observations are supported in the detection results figures, Fig. 10.

VI. CONCLUSION

This paper provides an idea on how deep learning based object detection can be utilized for construction safety measures. The work majorly concentrates on Faster R-CNN with Inception V2 as backbone architecture to implement the detection system. In out experiments, we are able to detect the helmets of the construction workers and the individuals without helmet with high confidence values of bounding boxes in Hard hat workers dataset. Yet the results obtained cannot be comparatively studied with other researches or projects as there was no study done on the algorithm with the current utilised dataset. The mAP values can be further increased by tuning more configuration parameters with a dedicated GPU based training. As per our analysis, if a real time detection via webcam feature has to be included for future directions, then the current backbone architecture would not be a good choice as in terms of speed there are faster backbone architectures available which can mostly overcome current backbone architecture's performance.

REFERENCES

[1] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu and Matti Pietik¨ainen, "Deep Learning for Generic Object Detection: A Survey" International Journal of Computer Vision (2020) 128:261–318, 31 October 2019

[2] A. Kelm, L. Laußat, A. Meins-Becker et al., "Mobile passive radio frequency identification (RFID) portal for automated and rapid control of personal protective equipment (PPE) on construction sites," Automation in Construction, vol. 36, pp. 38–52, 2013.

[3] S. Barro-Torres, T. M. Fern´andez-Caram´es, H. J. P´erez-Iglesias, and C. J. Escudero, "Real-time personal protective equipment monitoring system," Computer Communications, vol. 36, no. 1, pp. 42–50, 2012.

[4] A. H. M. Rubaiyat, T. T. Toma, M. Kalantari-Khandani et al., "Automatic detection of helmet uses for construction safety," in Proceedings of the 2016 IEEE ACM International Conference on Web Intelligence Workshops (WIW), ACM, Omaha, NE, USA, October 2016.

[5] K. Shrestha, P. P. Shrestha, D. Bajracharya, and E. A. Yfantis, "Hard-hat detection for construction safety visualization," Journal of Construction Engineering, vol. 2015, Article ID 721380, 8 pages, 2015.

[6] Fan Wu, Guoqing Jin, Mingyu Gao, Zhiwei HE, Yuxiang Yang, "Helmet Detection Based On Improved YOLO V3 Deep Model" Proceedings of the IEEE 16th International Conference on Networking, Sensing and Control, May 9-11 2019.

[7] Licheng Jiao, Fan Zhang, Fang Liu, Shuyuan Yang, Lingling Li, Zhixi Feng And Rong Qu, "A Survey of Deep Learning-Based Object Detection", IEEE ACCESS, VOL. 7, 2019

[8] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS, 2015.

[9] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision." In CVPR, 2016.

[10] https://public.roboflow.com/object-detection/hard-hat-workers/2

[11] Zhengxia Zou, Zhenwei Shi, Member, IEEE, Yuhong Guo, and Jieping Ye,"Object Detection in 20 Years: A Survey, arXiv:1905.05055v2 [cs.CV] 16 May 2019.

[12] Vipul Sharma, Roohie Naaz Mir ," A comprehensive and systematic look up into deep learning based object detection techniques: A review", ELSEVIER, Computer Science Review 38(2020) 100301 26 pages, Aug. 2020.