CONSTRUCTION SITE SAFETY MONITORING USING COMPUTER VISION

Ashwanth V
M.Tech, AI & DS
Muthoot Institute of Technology and Science
Kochi, India
mtech20ai03@mgits.ac.in

Ms.Dhanya Sudarsan
Asst. Professor
Muthoot Institute of Technology and Science
Kochi, India
dhanyasudarsan@mgits.ac.in

***Abstract*—The threat to worker safety and health is high in jobs such as construction work. As a safety measure, it is necessary to monitor the workers and ensure that they wear the right type of and adequate number of Personal Protective Equipment (PPE) at the building site. But keeping an eye on workers manually through CCTV footage would be a time-consuming operation. This paper introduces an approach to help monitor a construction site without the need for human interaction. Initially, YOLOv4 is used to detect construction workers, and the bounding boxes of each worker is cropped out into three halves. The cropped parts are then examined with EfficientNet to see if the specific PPE kit component is present. In addition, construction tools and equipment are identified, and a worker's safety score is determined based on the nearness of the worker to the discovered equipment.**

***Index Terms*—computer vision, yolov4, construction safety, efficientnetb5, transfer learning**

I. Introduction

Around 2.3 million workers are reportedly killed or seriously injured every year as a result of work-related accidents or illnesses according to a research conducted by the International Labor Organization (ILO). The mortality rate at workplaces is on the rise due to the lack of safety monitoring and precautionary measures followed. Trips, slips and falls, being hit by objects on head, getting caught in between machinery due to being in close proximity to machinery, etc., are some of the accidents that occur [30]. Being one of the most dangerous job sectors, construction is least digitized and yet incurs a huge demand for workers. Wearing proper safety gear and having good posture while working should be considered as safety measures. However, only a minimal number of workers tend to wear the PPE kits due to factors such as lack of awareness regarding the safety measures to be followed, discomfort to work while wearing these protective materials [21], carelessness, etc. Continuous monitoring is thus necessary to ensure that workers adhere to safety rules. Research has shown that the majority of the hazardous work-related injuries to the head can be avoided by wearing the hard-hats [31]. By absorbing shock from direct blows to the head, these helmets protect workers from skull fractures, neck sprains and concussions, caused by objects striking their heads or them falling from heights. Wearing the vest can help the workers locate each other from a distance or during situations with poor visibility. This can also help one understand the particular task a worker is assigned based on the colour of their vest and helmet. The type of shoes worn on the site is also important. Specially designed boots are the most common ones used. Traditionally, construction site supervisors or safety inspectors have been manually monitoring and enforcing safety restrictions on the workers in the construction sites. However, continuous supervision is difficult owing to human mistakes induced by distractions and through disproportion in the number of workers and supervisors.

In recent years, various methods based on wearable sensors such as RFID tags, UAVs for capturing construction images combined with deep learning algorithms, and various computer-vision based techniques have been proposed and implemented to aid in automating the process of construction safety monitoring. The introduction of deep neural networks has led to an improvement in object detection. But the majority of vision-based approaches simply focus on detecting the hard-hats and the workers as a whole, single process, lacking the accuracy of detection of the PPE kit components. Camera perspectives cause variations in the object scales, background complexities , etc., and add up to the difficulties of object detection. Existing methods usually link workers and PPE by determining whether the detected PPE is located in the worker's detection region or close to it. However, merely discovering this will not assist us in resolving the problem of safety. For example, a worker could be holding the hard-hat instead of wearing it, or the hard-hat could be placed on a table near to where the worker might be standing. In such cases, the existing methods would end up detecting the person and a hard-hat in the image/video and conclude that the worker has a hard-hat on, which is false. It is critical to determine whether the worker is wearing the hard hat on the head, if the worker is equipped with the full PPE kit, whether the worker is near any equipment of danger without the full kit, etc. Also, the CCTV that captured the video footage of a construction site could be at an angle that covers beyond the boundary of the construction area. In such cases, merely detecting the workers and their PPE will create false alarms. For example, a person who is outside the construction region, and not wearing the PPE since he could be just someone passing by the site, captured by the CCTV footage and recognized by the automation system, could create an alarm indicating a non-PPE kit wearing worker.

The majority of existing PPE kit detection only deals with hard hats. Multi-class PPE kit detection would definitely require a large collection of the individual PPE kit image dataset. There are multiple ways the problem of construction site safety monitoring can be approached. A construction specific dataset that includes construction machines, workers, and other equipment that is commonly observed on a construction site can be used to custom train an object detection or segmentation models. Object detection helps locate instances of certain objects in images or videos, but does not describe the shape of the objects, whereas segmentation would break down an image into various subgroups known as image segments and provide a granular understanding of the objects in the image. Algorithms like Region Based Convolutional Neural Networks (R-CNN), Fast R-CNN, Faster R-CNN, Single Shot Detector (SSD) and YOLO (You Only Look Once) are the most commonly used object detection algorithms, with YOLO being the most recent and having the highest accuracy amongst the others [32]. But segmentation techniques cannot distinguish the shading in the images when there is noise and high variations are present. Since construction site CCTV footage is of poor quality, the usage of object detection methods would be a better choice.

In the proposed approach, PPE kit dataset has been created from the existing large pool of construction site worker images, as will be discussed in section 3. For the main part, YOLOv4 has been used to detect the workers from the video and then split the bounding box of each of those identified workers into 3 halves. Using the method of transfer learning, three EfficientNet models trained on each of the PPE kit components, namely hard-hat, vest, and shoes, are then applied to the three cropped halves for each of the workers identified. Another YOLOV4 model that is custom trained to detect construction equipment and tools is also applied simultaneously to the video frames. A risk score is then determined based on the number of PPE kit components worn by the worker and their proximity to construction equipment, in order to determine the worker's safety. The approach has been able to solve the problem of verifying if the worker is wearing the PPE kits and also notify only the right people detected as workers using their safety score on the construction site.

II. Review of Literature

Recent years have seen a significant amount of attention being drawn towards the use of deep learning computer vision methods due to their self-learning capability of learning useful and important features. Some of the most popular convolutional neural networks for object detection and classification from images are Alex Nets, GoogLeNet and ResNet50 [1]. The most recent being the GoogLeNet and ResNet50 has better precision. Apart from the models or networks being used, importance has to be given to the dataset that is going to be used to train the model. When it comes to the field of construction, there has been multiple research conducted to collect image datasets from construction sites. In [37], the authors have used fully convolutional-based algorithms to detect worker's hard hats. In [2], the authors have created a benchmark dataset containing 15 different DNN-based detectors using the MOCS dataset. In [3], to validate the feasibility of ACID dataset, authors trained the data set using four existing deep learning object detection algorithms- YOLO-v3, Inception-SSD, R-FCN-ResNet101, and Faster-RCNN-ResNet101. The average detection speed of the four algorithms is 16.7 frames per second (fps), which satisfies the needs of most studies in the field of automation in construction. Authors in [4] have used the VGG-16 model, that is pre-trained on the ImageNet dataset, to investigate two categories of classification tasks namely, single-label classification and multi-label classification in construction imagery. Pictor v1.1 dataset has been used for the transfer learning of the VGG-16 model, and the authors have acquired an accuracy of 90% for single-label classification and 85% for multi-label classification. In [38], authors have collected recorded videos from tower crane camera and have performed image recognition task with Mask R-CNN method. Also, to identify the safety distance, they have used pixel and actual distance conversion method. Authors of [5] have presented three deep learning models built on YOLO-v3 architecture to verify PPE compliance of workers from images and videos in real-time. The model has been trained on the Pictor-v3 dataset and has achieved an average performance accuracy of 80%. First, hats and vests are detected, and a machine learning model, such as a decision tree, determines if each worker in the detected group is properly wearing hats or vests. The second method employs a convolutional neural network (CNN) framework to identify individual workers and verify their compliance with PPE. And the third approach first detects only the workers, and then assigns a classification score to each worker in the input image using CNN-based classifiers, based on the presence or absence of PPE kit. An improved version of SSD has been used in [13] to detect the safety helmets. The authors have used a fusion of multi-layer to identify low semantic and deep semantic information which helps in identifying the smaller targets [13]. In [7], authors have used YOLOv4 and YOLOv4-tiny for real-time detection of fire and PPE equipment at construction sites. The findings were utilised as a basis for proving the effectiveness of YOLOv4 algorithms in real-time detection and monitoring at construction sites [7]. Detecting whether the hard-hat is actually on the worker's head is yet another major challenge. Some methods include the authors using a MobileNet architecture to run the detector in real-time [6]. In this [6], a person's head is localized and then the MobileNet model is deployed to identify if the person is wearing hard-hat. In [36], workers and hard hats are detected separately but concurrently, and the amount of overlap between the worker's head and the detected hard helmets is used to determine whether or not the worker is wearing a hard hat. The authors of [8] have used a new approach to detect hard-hat wearers based on head key-point localization. Results show that it surpassed both, the solution based on the relative bounding box position of people and hard hats and the direct detection of hard hat wearers and non-

wearers. But the detection of head key-point is a difficult and error-filled process. Authors of [9] have used a pose estimator to detect worker body parts as spatial anchors and to localize the part attention regions. CNN based classifier is used to recognize both PPE and non-PPE classes within the attention regions. But the method uses a lightweight MobileNet rather than a deep network. Also, the implementation of the same on videos is not possible since pose estimation takes up a lot of resources.

III. Methodology

The Deep Learning models used here are YOLOv4 and EfficientNet-B5. YOLOv4 being an object detection model helps in identifying and representing the objects in the images and videos. EfficientNet-B5 on the other hand is an image classification algorithm that helps in classifying a set of images into a number of pre-defined classes.

A. BASE MODELS

1) EfficientNet-B5: EfficientNet is a variant of convolutional neural network. The authors of this model [9], studied model scaling and came to a conclusion that carefully balancing the depth, width and resolution of the CNN can help increase its performance. The researchers have used a novel approach called 'Compound Scaling' to scale the aforementioned dimensions. The Fig.1. shows the scaling methods comparisons as displayed in the original paper [9]. This compound

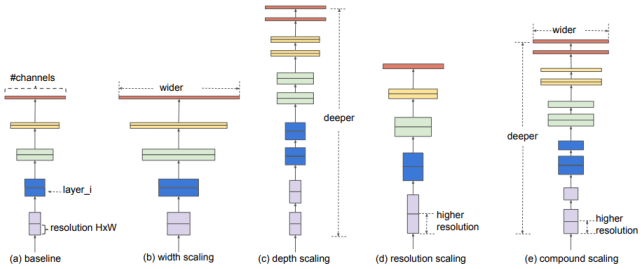

Fig. 1. EfficientNet Scaling

scaling method has eventually helped increase the accuracy and efficiency of models such as MobileNet(by +1.4%) and RestNet (by +0.7%) compared to the existing scaling methods. To create the baseline architecture, shown in Fig.2.

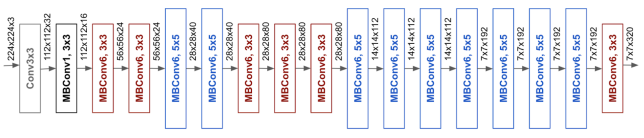

Fig. 2. EfficientNet Architecture

[9], neural architecture search was done using AutoML MNAS framework, which optimized both accuracy and efficiency. The resulting architecture uses mobile inverted bottleneck convolution, similar to MobilNetV2 [16]. EfficieNet's performance when compared with other models that was trained on ImageNet dataset [27] showed that the latest version, that is EfficientNet-B7, has the highest accuracy among all with a smaller number of parameters. ImageNet is a large dataset consisting of human annotated images that is intended for computer vision related studies and development [39]. There are more than fourteen million images in the dataset with about twenty one thousand classes. The EfficientNets were also tested on other datasets to check their performance on transfer learning tasks, and has achieved state-of-the-art accuracy on datasets such as CIFAR-100 [16]. Since Keras framework does not support the ImageNet weights for version B6 and B7, performing transfer learning on the versions is not possible. Hence the version B5 is used in this paper.

2) YOLOv4: YOLOv4 is an improvement to the YOLOv3 algorithm by roughly 10% in mean average precision and 12% in frames per second.

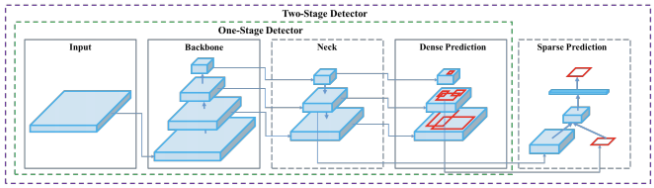

Fig. 3. YOLOv4 Architecture

The architecture, shown in Fig.3. [15], has four distinct blocks: backbone, neck, dense prediction and sparse prediction. The backbone helps in feature extraction and uses CSP-Darknet53. The Neck works similar to a ResNet. It also uses a modified Path aggregation network, that helps aggregate the information to improve accuracy. The head helps locating the bounding boxes and for classification. Specific to YOLOv4 is the feature called bag of freebies that helps improve accuracy during training without increasing inference time [15]. Also, a method called bag of specials is being used, that slightly increases inference cost but helps significantly improve the accuracy of object detection. The activation function used in YOLOv4 is called Mish activation function. It is a novel self-regularized non-monotonic activation function whose formula is defined below [15].

$$f(x) = xtanh(softplus(x)) \tag{1}$$

The Fig.4. shows the Mish activation function graph in comparison to other activation functions. The Mish activation function has helped solved the Dying ReLU phenomenon due to the preservation of a small amount of negative information. YOLOv4 is trained and tested on the COCO dataset that contains around 80 object classes. YOLOv4 outperforms other object detection models relative to inference speed and also makes it easier to create custom object detection tasks with high performance rates [40].

B. BASE APPROACH

The workflow diagram of the overall methodology is shown in Fig.5. Initially the video is provided as input, which then

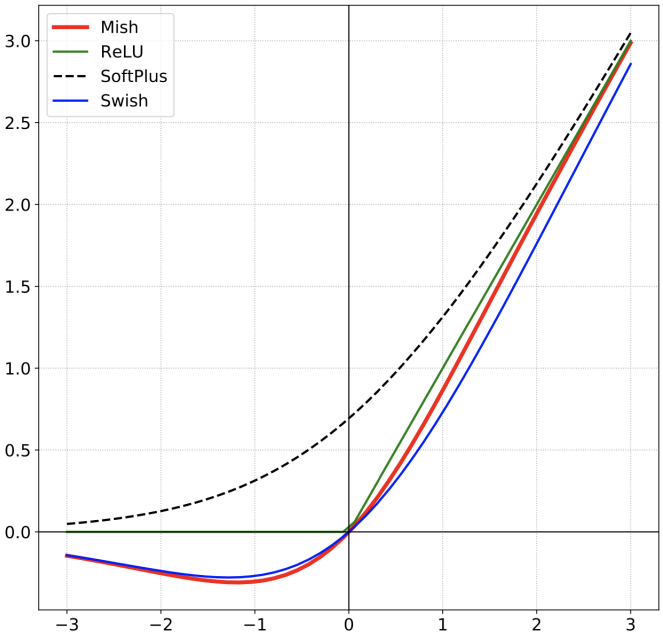

Fig. 4. YOLOv4 activation functions graph

is converted into frames for further tasks. One part performs person detection using YOLOv4 and the other performs construction equipment detection using an YOLOv4 model that is custom trained through transfer learning process. The detected

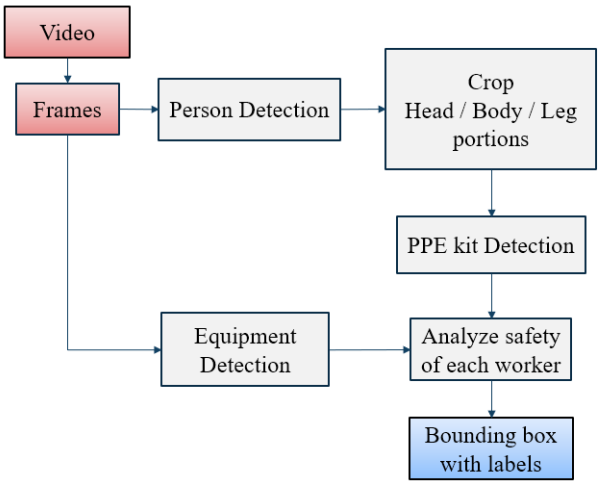

Fig. 5. Work Flow Diagram

person's bounding boxes are then divided into three halves. The EfficientNet models that were trained on each of the PPE kit components (Hard-Hat, Vest, and shoes) are then applied to the divided regions. The output of the PPE kit detection along with the construction equipment detection is used to calculate a safety score. Based on this safety score, the workers are marked as safe or in danger.

C. DATASET CREATION

1) PPE Kit: Publicly available construction site images such as CHV dataset, Pictor v3, and other web mined images have been used. The CHV dataset was created by Wang,et.al. [17], and it consists of 1330 high quality images from real construction sites with consideration given to different construction locations, gestures, varied angles and distances. The Pictor-v3 dataset [5] consists of 774 crowd-sourced and 698 web-mined images, and contains 2,496 and 2,230 instances of workers, respectively. But these collections of images consist of a variety of complexities within each of the image, has to be sorted for our use case. For example, an image could contain a total of 5 people out of which three could be wearing a hard-hat and a vest but no shoe, one wearing only a hard-hat, and another wearing none of the PPE kit components. Such images cannot not be used for detecting individual PPE kits since having multiple cases within an image could meddle with accuracy. Moreover, detecting individual persons is the key to the success of the overall system and thus single worker images have to be used for training. The availability of such images of individual workers or individual PPE kit components is limited or non-existent. Thus, an approach of extracting only the necessary objects from images for further training of models is used here. The procedure starts with the

Fig. 6. Image cropping for dataset creation

use of a basic YOLOv4 model that is fed with PPE kit dataset images to identify workers. YOLOV4 is trained on COCO dataset that consists of 'person' as an object class. Thus, a pre-trained YOLOv4 model is being used here to identify individual workers in the images. The identified workers are represented using a bounding box, which is how a one-stage

detector like YOLO works. The bounding box of the individual workers are divided into three portions from the top. The division is performed by a simple method of splitting the whole image into four quarters. The first quarter or the $1/4^{th}$ part is chosen as the head region, the middle halves or the $2/4^{th}$ and $3/4^{th}$ part is chosen as body region and the last quarter is chosen as the leg region. A sample working of image cropping and splitting method explained above is demonstrated in Fig.6. The cropped images are saved into separate files as head, body and leg. Due to the lack of clarity of images, or complexities like occlusion, the workers detected or the images cropped could be of the wrong type that cannot be used for our use-case. For example, a person bending down detected by the model, when cropped, the chest portion of the person ends up being put into the head portion folder. Or situations where two people standing behind each other, gets detected by the model as a single person, and thus when cropped creates multiple heads in a single image for the hard-hat detection folder. These saved cropped images are thus manually sorted into different classes, such as, Hard-hat / No-hard-hat, Vest / No-vest, and Shoes / No-shoes. The images are then used for the training of EfficientNet-B5 models.

2) Construction Equipment: Datasets such as Alberta Construction Image Dataset (ACID) [3] and Moving Objects in Construction Sites (MOCS) [2] along with some web-mined images have been used for construction equipment / machinery detection. Open Images v6 tools dataset and other web mined tools dataset have been used for tools detection. ACID dataset contains 2,850 construction images with 6,500 machine objects of 3 different types, namely, excavator, dump truck and concrete mixer truck. The images of the ACID dataset are collected through online collection (web crawler) and onsite collection (UAVs, CCTV cams, manual clicks) [3]. The MOCS dataset contains 41,668 images collected from 174 different construction sites and thirteen different categories of moving objects are present. These datasets have been pre-labelled but consists of multiple classes. Since we are combining all the images together, and also having multiple classes could create class imbalance problem, the images are being custom labelled. The original set of construction machinery images containing objects such as different types of trucks, JCBs, cranes, etc., are all labeled under a single label 'equipment'. And the tools images consisting of hammers, screwdrivers, welding tool, etc., are labelled under a single label 'tool'. CVAT [14] software has been used for the custom labelling process. CVAT helps us create the bounding boxes for object detection tasks and export the custom labeled images in any preferred format. The format used here is YOLO, which consists of images and its corresponding bounding box coordinates as a text file.

D. Transfer Learning

To attain a high-performance rate, training a model from scratch is not only computationally expensive, but also necessitates a substantial amount of data. Transfer learning is a process in machine learning where a pre-trained model is reused to perform a new task [33]. The knowledge that the model has on the previous assignments it performed helps increase the prediction accuracy on the new task in less amount of training time and training data. This has been proved by studies performed by authors of [37]. Transfer learning can be performed on similar or different datasets in terms of its pre-trained dataset classes. That is, a model that is trained on dog images, retrained using cat images is considered as similar set of data since both cats and dogs have eyes or other such similar features. Whereas, images of plants being trained on a model that was pre-trained on dog images would be considered to be different set of data. In our paper, the transfer learning is being performed on similar set of data for both YOLOv4 and EfficientNet-B5 models. The major steps in the process of transfer learning includes, removing the ends of the fully connected neural network, replacing or adding a new fully-connected layer with output neuron dimension equal to the number of classes in the dataset used for transfer learning, freezing all the weights of the initial set of layers, and then training the network to update the weights. Unfreezing the layers and updating the weights after the initial training can be done for tuning the model for further accuracy improvements if any.

E. Data Augmentation

Machine learning models require large amount of data for its training. Data augmentation is a method which helps improve the performance and outcomes of any machine learning models by creating new and different samples from the existing training dataset [34]. Types of augmentation techniques include padding, random rotation, re-scaling, vertical and horizontal flipping, gray-scaling, etc. Horizontal or vertical flipping re-positions pixels while preserving image features. Rotation augmentations are done by rotating the image right or left on an axis between 1°and 359° [**?**]. Too much increase in the rotation degree can cause the label to be no longer preserved to the image.

F. PPE Kit training

EfficientNet-B5 that is pre-trained on ImageNet dataset is used here. The ImageNet dataset consists human annotated images of varied classes. Transfer learning is implemented on the model and helps reduce the training time and the accuracy.

Each of the PPE kit component dataset is given as input to three different EfficientNet B5 models. Fig.7. shows the model training process of PPE kit. By default, the input image size to be given for EfficientNet B5 model is 456 x 456 x 3. Data augmentation of random horizontal flipping and random rotation of 20°is also performed as pre-processing apart from image re-sizing. The parameters are set as, dropout rate as '0.2', learning rate as '1e-2', and the loss function used is 'categorical cross entropy'. Categorical cross entropy is well suited for classification tasks and helps distinguish two discrete

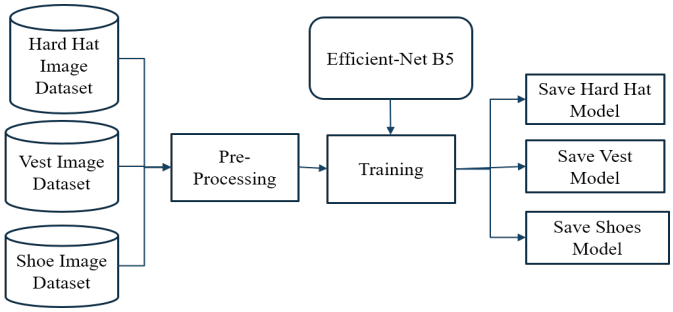

Fig. 7. Flow diagram of PPE Kit models training

probability distributions from each other [41]. The formula of the loss function is:

$$Loss = -\sum_{i=1}^{outputsize} y_i.log\hat{y}_i = 1 \quad (2)$$

where, $\hat{y}_i$ is the i-th scalar value in the model output, y_i is the corresponding target value and output size is the number of scalar values in the output [41]. Learning rate helps decide how much gradient is to be back propagated. The learning rates were tested from '1e-1' to '1e-5', and '1e-2' provided the better result. Each of the model is trained for 50 epochs and the weights are saved. Hyper-tuning was not necessary since the initial training provided a high accuracy rate. Nevertheless, the models were further trained for 50 more epochs, adding up to a total of 100 epochs, but the accuracy either didn't have a change or had a decrease. Hence the final model weights used are the ones obtained in the 50-epoch training process.

G. Construction Equipment Training

YOLOv4 is a one-stage object detection model that has been widely used for object detection tasks due to its significantly higher accuracy and detection speed than majority of the other detection models such as SSD, Faster R-CNN or even older versions of the YOLO family. The custom labelled images are fed into the pre-trained YOLOv4 model to perform transfer learning.

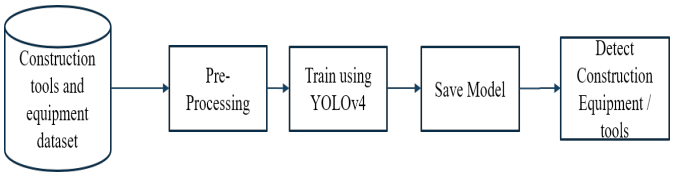

Fig. 8. Flow diagram of equipment detection model training

The Fig.8. shows the custom training steps of YOLOv4. The images are resized into 416 × 416 × 3, which is the default input image size that is needed for YOLO as input. Since transfer learning is being performed, the parameters of the YOLOv4 model have to be changed. The batch size is set as 64 and subdivisions as 16. Batch size is the number of training examples used in one iteration. Using a larger batch size will lead to poor generalization, but using a very small batch size doesn't guarantee to converge to the global optima. The height and width are set as 416x416. Max_batch is calculated as (number of classes) *2000. But this number should not be lesser than 6000, thus if the number of classes is 1,2 or 3, the value should be 6000, however, for 5 classes it would be 10000. Steps is set as 80% of max_batches and 90% of max_batches. So, if max_batch is 10000, then steps=8000,9000. Filters is set as equal to the (number of classes +5) * 3. So, for example, if there is one class, then filters is equal to 18. The model is trained for 6000 iterations and the model weight is saved.

H. Safety Detection

As seen in Fig.9., the system consists of two. The video given as input to the system, which will be converted to frames, is passed as input to both the modules. In the first module, YOLOv4 will be used to detect the person and bounding boxes of the detected persons will be split into two halves horizontally, obtaining 3 regions. The splitting is done similar to the one performed during dataset creation explained in Section 3c. The custom trained EfficientNet models are then

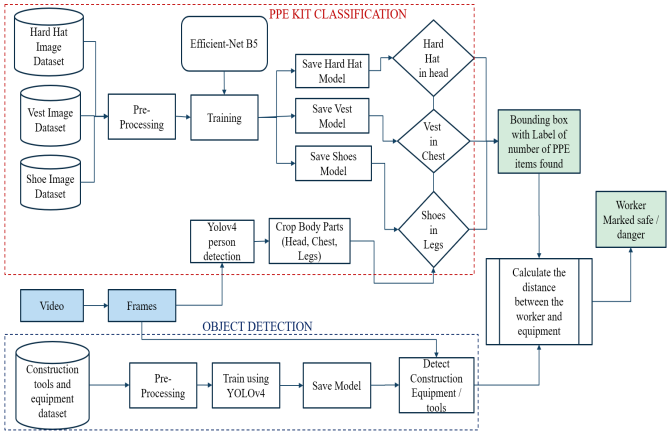

Fig. 9. Overall System Architecture

applied to each of the split regions, i.e., the hard-hat detection model is applied to the head region, the vest detection model is applied to the middle/body region and the shoe detection model is applied to the lower/leg region. The output of the module will consist of a label and bounding box over the worker, where the label represents the number of PPE kit components present. The number of PPE kit component ranges from 0 to 3, with '0' representing that no PPE kit has been detected and '3' representing that all the 3 PPE kit has been detected on a particular person. In the object detection module, custom trained YOLOv4 model will check for the construction equipment, i.e., machinery and tools, in the video frame simultaneously when the first module is running. The identified objects will be represented with a bounding box and a label with the detected class label name. The outputs of the first and the second module will be combined to perform further tasks of the system. A person identified through the first module, who is not wearing any of the component will

be marked as in 'Danger zone' and the person wearing at least one component will be marked as 'Safe zone' by default. If a person who is not wearing any of the three PPE kit items, is standing near an equipment, then that particular person will be marked in red bounding box with a label 'Equipment Danger'. The distance between the person and the equipment is calculated using Euclidean distance 'd' of the centers of the two objects given by:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{3}$$

A threshold distance of 250 is set as default, and if the obtained 'd' is lesser than the threshold, then the objects are considered to be near. This would significantly solve the problem of detecting a person outside of the construction zone without PPE kit and falsely notifying the authorities.

IV. Results Obtained

After the dataset creation process, a total of 574 of 'hard-hat', 210 of 'no_harhat', 825 of 'vest', 278 of 'no_vest', 192 of 'shoes' and 170 of 'no_shoes' images were obtained. These images are then split as training set, validation set and test set in the ratio of 8:1:1.

A. PPE kit detection

From the EfficientNet-B5 model used for PPE kit classification, we have obtained an accuracy of 98.48% for hard-hat, 99.43% for vest and 91.00% for shoes.

1) Hard-Hat: The initial loss and accuracy of the pre-trained EfficientNet-B5 model was 0.65 and 56.9% respectively.

After performing transfer learning on the model with hard-hat

Fig. 10. hard-hat Classification Output

images, the evaluation obtained are:
loss = 0.0874, accuracy = 0.9848, validation_accuracy = 0.9861 and test_accuracy = 0.9850. Fig.10. shows the sample classification output obtained after training the hard-hat EfficientNet model.

Model performance curves obtained are shown in Fig.10.

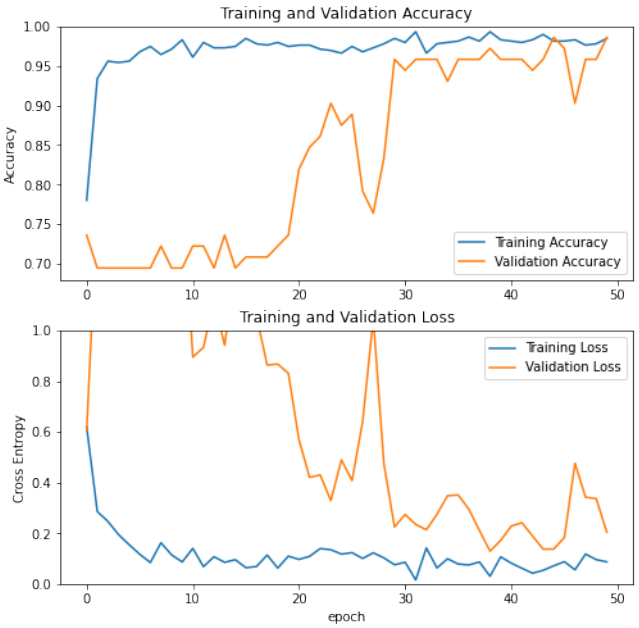

Fig. 11. Hard-Hat Model Training Performance Graph

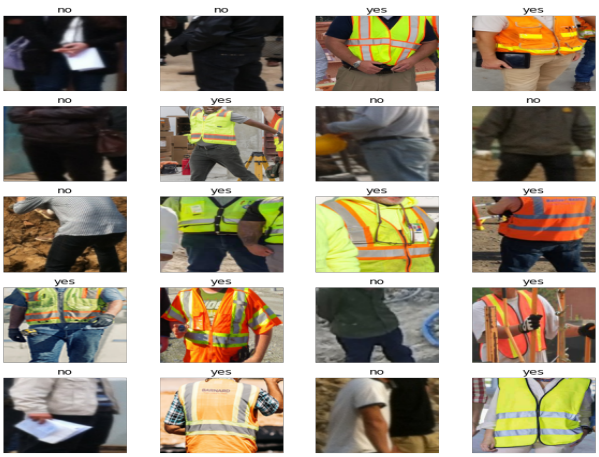

Fig. 12. Vest Classification Output

2) Vest: The initial loss and accuracy of the pre-trained EfficientNet-B5 model was 0.66 and 61.5% respectively. After performing transfer learning on the model with vest images, the evaluation obtained are:
loss = 0.0357, accuracy = 0.9943, validation_loss = 0.0910, validation_accuracy = 0.9817 and test_accuracy = 0.9897. Fig.12. shows the sample classification output obtained after training the hard-hat EfficientNet model. Model performance curves obtained for the vest model is shown in Fig.13.

3) Shoes: The initial loss and accuracy of the pretrained EfficientNetB5 model was 0.72 and 50.0% respectively. After performing transfer learning on the model with vest images, the evaluation obtained are:
loss = 0.03932, accuracy = 0.9100, validation_loss = 0.5848, validation_accuracy = 0.7778 and test_accuracy = 0.8157.

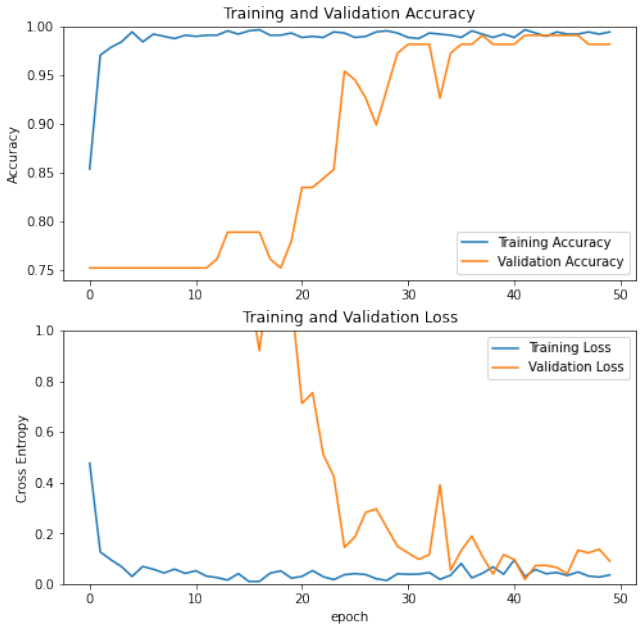

Fig. 13. Vest Model Training Performance Graph

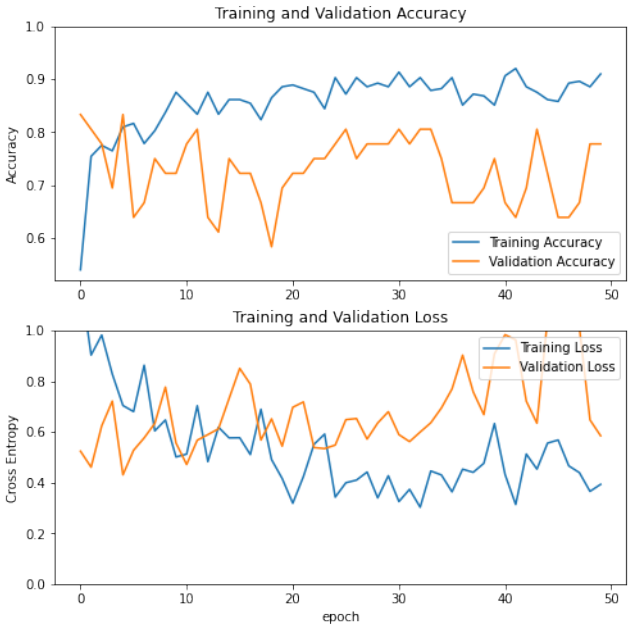

Fig. 15. Shoes Model Training Performance Graph

Fig. 14. Shoes Classification Output

Fig.14. shows the sample classification output obtained after training the shoes EfficientNet model. The model performance curves obtained for shoes model is shown in Fig.15.

B. Equipment Detection

About 3000 images of construction equipment and 100 images of tools were used for the training. An Average Precision of 73.81% has been obtained for equipment detection and 38.05% for tools detection.

$$Precision = TP/(TP + FP) \tag{4}$$

where, TP is True positive and FP is false positive.

$$Recall = TP/(TP + FN) \tag{5}$$

where, FN is False negative.

$$F1-score = (2*Precision*Recall)/(Precision+Recall) \tag{6}$$

The overall model has a Precision of 0.63, Recall of 0.75 and F1-score of 0.68. And the mean average precision of the model is 55.93%.

Name	AP(%)	TP	FP	Precision
Equipment	73.81	655	371	0.63
Tool	38.05	29	31	0.48
Overall Model	55.9	684	402	0.63

Table.1. shows the model evaluation,i.e, the performance of the custom trained YOLOv4 model after the 6000-iteration training process. The chart in Fig.16. shows the average loss vs iteration. For a model to be accurate, a loss under 2 should be aimed for. Fig.17 and Fig.18 shows the equipment detection output where the Trucks and JCBs are detected. Fig.19 and Fig.20 shows a sample of tools being detected using the custom trained YOLOv4 model.

C. Final Output

The overall system was tested on multiple videos that wear obtained on public websites. Figures Fig.21. and Fig.22. shows some of the screenshots from the output. In Fig.21, the worker is seen wearing shoes and vest and no hard-hat. Hence, the label shows 'Safety:2' for that worker. Also, since the worker is safe from any equipment danger, he is labelled as in 'Safe zone'. Whereas in Fig.22, there is a construction equipment identified and also a person near it. Since there is an equipment present, the person is labelled as in a 'Danger zone'. The worker is seen wearing only a hard-hat and not any of the remaining set of PPE kit items, thus, he is marked with a label

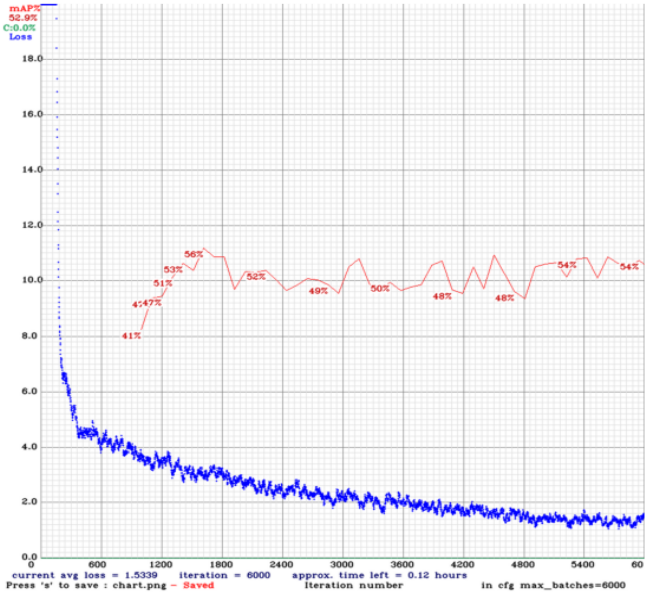

Fig. 16. YOLOv4 training performance graph

Fig. 17. Construction Equipment Detection 1

Fig. 18. Construction Equipment Detection 2

Fig. 19. Tool Detection 1

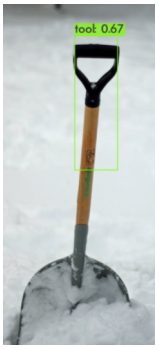

Fig. 20. Tool Detection 2

Fig. 21. Safety Detection Video Output 1

Fig. 22. Safety Detection Video Output 2

'Safety:1'. Since he is not equipped with all the three PPE kit components and also is near the construction equipment,i.e., the distance between the centroids of the bounding boxes of the worker and the equipment is lesser than the set threshold, he is labeled as to be in an 'Equipment Danger'.

V. Discussion

The dataset creation approach has provided us with the proper set of images that could be used for our use-case of having to train individual PPE kit items. The number of images is less because of choosing only the best and high quality images from the cropped set, and because of using the best set of images, the accuracy of EfficientNet models are at the highest rate when compared to the existing methods. The learning rates for the Hard-hat and Vest model have been stable, i.e., there was not much learning happening, after the 30th epoch. The YOLOv4 model that was used for construction machinery and tool detection has provided an average accuracy in comparison to industry standards. This was probably because of inaccuracies in the custom bounding box labeling process. Nevertheless, the model has been able to identify the objects correctly with slight inaccuracy in the placement of bounding boxes over the objects, which can be neglected in our use-case. The final model has been able to correctly recognize the worker and the equipment in a single frame, and also provide the safety of the worker. Since the available CCTV footage that can be used for testing are of less quality, there are inaccuracies in the output in certain frames of the video.

VI. Conclusion and Future work

This work has been implemented to solve the problem of the safety of workers on construction sites by identifying if the workers are wearing the PPE kit components and also their nearness to construction equipment. The PPE kit items recognized in this paper are Hard-hat, Vest and Shoes. The existing methods had the disadvantage of being able to only identify the worker and PPE kit as a whole, and thus reducing the accuracy, or only being able to implement the system for single PPE kit detection. Using the EfficientNet framework, which is a fast and accurate image classification model, and YOLOv4, we have been able to provide a fast and accurate automated safety monitoring system. Creating a dataset with different PPE kit components from whole images, and applying a custom trained model onto the different body parts of workers has helped solve the problem of verifying the placement of PPE on workers. Also, having created a safety score for the identified persons by considering whether the worker is near an equipment or not, we have been able to solve the false notifications of persons that are on-site and off-site.

Though the method works well, there are a few advancements that could be done to further improve the monitoring system. The quality of the images used to train the EfficientNet classifier could be improved by adopting super resolution (SR) techniques. The output has the problem of occlusion in some frames, which could be improved probably by detecting the workers using techniques like pose estimation at the cost of FPS and other resources. More types of PPE components could be detected, such as gloves, goggles, and also domain-specific types of PPE kits can also be detected.

References

[1] Neha Sharma, Vibhor Jain, Anju Mishra, An Analysis Of Convolutional Neural Networks For Image Classification, Procedia Computer Science, Volume 132, 2018, Pages 377-384, ISSN 1877- 0509

[2] An Xuehui, Zhou Li, Liu Zuguang, Wang Chengzhi, Li Pengfei, Li Zhiwei, Dataset and benchmark for detecting moving objects in construction sites, Automation in Construction, Volume 122, 2021, 103482, ISSN 0926-5805.

[3] Bo Xiao, Shih-Chung Kang, Development of an Image Data Set of Construction Machines for Deep Learning Object Detection, Journal Article, 2021, Journal of Computing in Civil Engineering, 05020005, 35, 2, 10.1061/(ASCE)CP.1943-5487.0000945.

[4] Nath N D, Chaspari T, Behzadan A H (2019). Single- and multi-label classification of construction objects using deep transfer learning methods, ITcon Vol. 24, Special issue Virtual, Augmented and Mixed: New Realities in Construction, pg. 511-526, doi:https://www.itcon.org/2019/28

[5] Nipun D. Nath, Amir H. Behzadan, Stephanie G. Paal, Deep learning for site safety: Real-time detection of personal protective equipment, Automation in Construction, Volume 112, 2020, 103085, ISSN 0926-5805,doi: https://doi.org/10.1016/j.autcon.2020.103085

[6] Wang, Lu, Liangbin Xie, Peiyu Yang, Qingxu Deng, Shuo Du, and Lisheng Xu. 2020. "hard-hat-Wearing Detection Based on a Lightweight Convolutional Neural Network with Multi-Scale Features and a Top-Down Module" Sensors 20, no. 7: 1868. doi: https://doi.org/10.3390/s20071868

[7] Kumar, S., Gupta, H., Yadav, D. et al. YOLOv4 algorithm for the real-time detection of fire and personal protective equipments at construction sites. Multimed Tools Appl (2021). doi:https://doi.org/10.1007/s11042-021-11280-6

[8] Ruoxin Xiong, Pingbo Tang, Pose guided anchoring for detecting proper use of personal protective equipment, Automation in Construction, Volume 130, 2021, 103828, ISSN 0926-5805, https://doi.org/10.1016/j.autcon.2021.103828

[9] Tan, M., Le, Q.V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. ArXiv, abs/1905.11946.

[10] Nath Nipun D., Behzadan Amir H, "Deep Convolutional Networks for Construction Object Detection Under Different Visual Conditions", Frontiers in Built Environment, volume 6, 2020.

[11] Yu Fan and Xianqiao Chen and Jinguang Xie and Zhenjie Fu, "An Algorithm for Detecting the Integrity of Outer Frame Protection Net on Construction Site Based on Improved SSD", Journal of Physics: Conference Series, IOP Publishing, 2021.

[12] Jixiu Wu, Nian Cai, Wenjie Chen, Huiheng Wang, Guotian Wang, "Automatic detection of hard-hats worn by construction personnel: A deep learning approach and benchmark dataset", Automation in Construction, Volume 106, 2019.

[13] Bin Dai, Yuhu Nie, Wenpeng Cui, Rui Liu, and Zhe Zheng, "Real-time Safety Helmet Detection System based on Improved SSD". In Proceedings of the 2nd International Conference on Artificial Intelligence and Advanced Manufacture AIAM2020.

[14] Boris Sekachev,et.al.,opencv/cvat: v1.1.0,aug,2020,Zenodo,v1.1.0,10.5281/zenodo.4009388. https://doi.org/10.5281/zenodo.4009388.

[15] Bochkovskiy, Alexey, Chien-Yao Wang, and Hong-Yuan Mark Liao. "Yolov4: Optimal speed and accuracy of object detection." arXiv preprint arXiv:2004.10934 (2020).

[16] https://ai.googleblog.com/2019/05/efficientnet-improving-accuracy-and.html.

[17] Wang, Z.; Wu, Y.; Yang, L.; Thirunavukarasu, A.; Evison, C.; Zhao, Y. Fast Personal Protective Equipment Detection for Real Construction Sites Using Deep Learning Approaches. Sensors 2021, 21, 3478. https://doi.org/10.3390/s21103478.

[18] Centers for Disease Control and Prevention, Traumatic brain injuries in construction, https://blogs.cdc.gov/niosh-science-blog/2016/03/21/constructiontbi/ , Accessed date: 6 July 2019.

[19] OSHA, Safety vest requirements to protect flaggers from traffic hazards during construction work, https://www.osha.gov/laws-regs/standardinterpretations/ 2002-03-11 , Accessed date: 6 July 2019.
[20] OSHA, Safety and health regulations for construction, https://www.osha.gov/lawsregs/regulations/standardnumber/1926/1926.28 , Accessed date: 6 July 2019.
[21] Akbar-Khanzadeh F. Factors contributing to discomfort or dissatisfaction as a result of wearing personal protective equipment. J Hum Ergol (Tokyo). 1998 Dec;27(1-2):70-5. PMID: 11579702.
[22] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2017) 1137–1149, https://doi.org/10.1109/tpami.2016.2577031.
[23] R. Girshick, Fast R-CNN, Proc. IEEE International Conference on Computer Vision, Santiago, Chile, 2015, pp. 1440–1448, https://doi.org/10.1109/iccv.2015.169.
[24] K. He, G. Gkioxari, P. Dollar, R. Girshick, Mask R-CNN, Proc. IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2961–2969, , https://doi.org/10.1109/iccv.2017.322.
[25] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, A.C. Berg, SSD: Single shot multibox detector, Proc. European Conference on Computer Vision, Amsterdam, the Netherlands, 2016, pp. 21–37, https://doi.org/10.1007/978-3-319-46448-0_2.
[26] M. Oquab, L. Bottou, I. Laptev, J. Sivic, Learning and transferring mid-level image representations using convolutional neural networks, Proc. IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, 2014, pp. 1717–1724, , https://doi.org/10.1109/cvpr.2014.222
[27] H.C. Shin, H.R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, R.M. Summers, Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning, IEEE Trans. Med. Imaging 35 (5) (2016) 1285–1298, https://doi.org/10.1109/tmi.2016.2528162.
[28] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, Proc. IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, 2009, pp. 248–255, , https://doi.org/10.1109/cvpr.2009.5206848.
[29] M. Everingham, L.V. Gool, C.K.I. Williams, J. Winn, A. Zisserman, The pascal visual object classes (VOC) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338, https://doi.org/10.1007/s11263-009-0275-4.
[30] Bureau of Labor Statistics, Fatal occupational injuries counts and rates by selected industries, https://www.bls.gov/news.release/cfoi.t04.htm , Accessed date: 6 July 2019.
[31] X. Huang, J. Hinze, Analysis of construction worker fall accidents, J. Constr. Eng. Manag. 129 (3) (2003) 262–271, https://doi.org/10.1061/(asce)0733-9364(2003)129:3(262).
[32] Srivastava, S., Divekar, A.V., Anilkumar, C. et al. Comparative analysis of deep learning image detection algorithms. J Big Data 8, 66 (2021). https://doi.org/10.1186/s40537-021-00434-w.
[33] Zhuang, Fuzhen, et al. "A comprehensive survey on transfer learning." Proceedings of the IEEE 109.1 (2020): 43-76.
[34] P. Kaur, B. S. Khehra and E. B. S. Mavi, "Data Augmentation for Object Detection: A Review," 2021 IEEE International Midwest Symposium on Circuits and Systems (MWSCAS), 2021, pp. 537-543, doi: 10.1109/MWSCAS47672.2021.9531849.
[35] Shorten, C., Khoshgoftaar, T.M. A survey on Image Data Augmentation for Deep Learning. J Big Data 6, 60 (2019). https://doi.org/10.1186/s40537-019-0197-0.
[36] Z. Xie, H. Liu, Z. Li, Y. He, A convolutional neural network based approach towards real-time hard hat detection, Proc. IEEE International Conference on Progress in Informatics and Computing (PIC), IEEE, 2018, pp. 430–434, , https://doi.org/10.1109/pic.2018.8706269.
[37] Oquab M., Bottou L., Laptev I. and Sivic J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. Proceedings of IEEE conference on computer vision and pattern recognition. 1717-1724.
[38] Yang, Zhen, Yongbo Yuan, Mingyuan Zhang, Xuefeng Zhao, Yang Zhang, and Boquan Tian, "Safety Distance Identification for Crane Drivers Based on Mask R-CNN" Sensors 19, no. 12: 2789 . 2019.
[39] J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.
[40] J. -a. Kim, J. -Y. Sung and S. -h. Park, "Comparison of Faster-RCNN, YOLO, and SSD for Real-Time Vehicle Type Recognition," 2020 IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia), 2020, pp. 1-4, doi: 10.1109/ICCE-Asia49877.2020.9277040.
[41] Zhilu Zhang and Mert R. Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. In Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18). Curran Associates Inc., Red Hook, NY, USA, 8792–8802.