

ORIGINAL ARTICLE

Journal Section

An IoT-based Human Detection System for  
Complex Industrial Environment with Deep  
Learning Architectures and Transfer Learning

Imran Ahmed PhD<sup>1\*</sup> | Marco Anisetti PhD<sup>2†</sup> |  
Gwanggil Jeon PhD<sup>3‡</sup>

<sup>1</sup>Center of Excellence in Information  
Technology, Institute Of Management  
Sciences, 1-A, Sector E-5, Phase VII,  
Hayatabad, Peshawar, KPK, Pakistan.  
E-mail: imran.ahmed@imsiences.edu.pk  
<sup>2</sup>Dipartimento di Informatica (DI),  
Università degli Studi di Milano, Via Celoria  
18, Milano (MI) 20133, Italy. E-mai:  
marco.anisetti@unimi.it  
<sup>3</sup>Department of Embedded Systems  
Engineering, Incheon National University,  
Incheon. Korea. E-mail: gjeon@inu.ac.kr

**Correspondence**  
Gwanggil Jeon PhD, Department of  
Embedded Systems Engineering, Incheon  
National University, Incheon, Korea  
Email: gjeon@inu.ac.kr

Funding information

Artificial Intelligence (AI), combined with the Internet of things (IoT), plays a beneficial role in various fields, including intelligent surveillance applications. With IoT and 5G advancement, intelligent sensors and devices in the surveillance environment collect large amounts of data in the form of videos and images. These collected data require intelligent information processing solutions, help analyze the recorded videos and images to detect and identify various objects in the scene, particularly humans. In this work, an automated human detection system is presented for a complex industrial environment, in which people are monitored/ detected from a top view perspective. A top view is usually preferred because it can provide sufficient coverage and enough visibility of a scene. This work demonstrates the applications, efficiency, and effectiveness of deep learning architectures, i.e., Faster-RCNN, SSD, and YOLOv3, with transfer learning. Experimental results reveal that with additional training and transfer learning, the performance of all detection architectures is significantly improved. The detection results are

**Abbreviations:** ABC, a black cat; DEF, doesn't ever fret; GHI, goes home immediately.

\* Equally contributing authors.

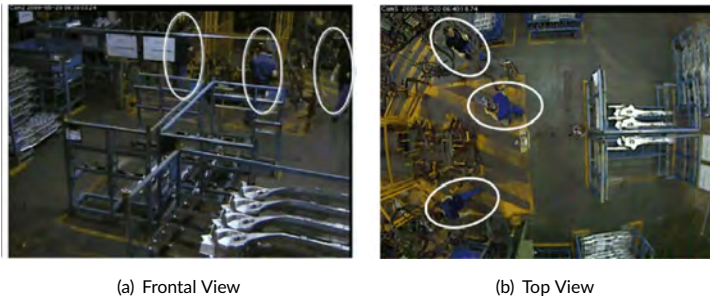
also compared using the same data set. The deep learning architectures achieve promising results with maximum True Positive Rate (TPR) of 93%, 94%, and 94% for Faster-RCNN, SSD, and YOLOv3, respectively. Furthermore, a detailed study is performed on output results that highlight challenges and probable future trends.

#### KEYWORDS

Internet of Things (IoT), Artificial Intelligence, Complex Industrial Environment, Person Detection, Top View, Deep Learning

## 1 | INTRODUCTION

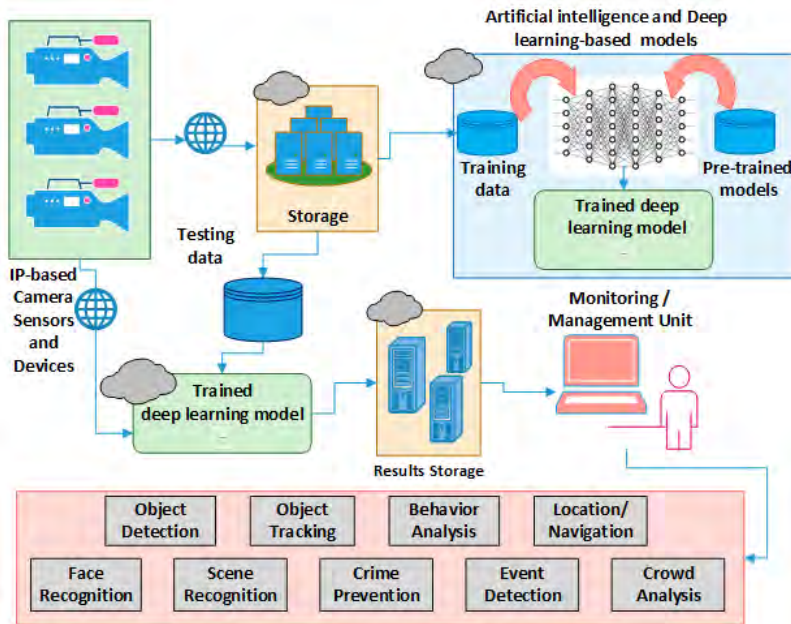
Intelligent systems have been widely utilized for information processing in modern society due to the rapid growth of artificial intelligence applications. Artificial Intelligence and the Internet of Things has recently developed as one of the hot research topics; it merges Artificial intelligence technologies with the Internet of Things infrastructure and offers more effective IoT service. It has shown remarkable success for various surveillance applications in different indoor and outdoor complex industrial environments. The abrupt changes in motion, cluttered scenes, camera viewpoints, a close interaction between individuals, and occlusion are crucial factors that might influence the performance of human detection algorithms [1]. The person's visual appearance significantly changes in terms of the body's rotation, poses, articulations, movements, scales, sizes, height, and torso, as can also be seen in FIGURE 1. The primary concern in the frontal view perspective is that it might suffer from occlusion (phenomena happen when the human body is completely or partially covered with other object or person), as presented in FIGURE 1(a).



**FIGURE 1** Sample images of the same complex industrial environment captured at the same time using two distinct camera perspectives and positions (a) White ellipses show people are occluded with heavy machinery in frontal view (b) People can be easily seen from the top perspective as shown in white ellipses.

It can be noticed in white ellipses that the people in the industrial environment are obscure with machines. To resolve this problem, some researchers, e.g., [2] proposed to utilize a single top-view camera as it provides better coverage and also helps to overcome occlusion problems. From sample images of FIGURE 1(b), it can be observed that utilizing a top view camera reduces occlusion problems. Along with resolving occlusion issues, it also reduces privacy issues

and might overcome power consumption, human resource, and installation expenses [3]. FIGURE 1 reflects the central contrast between two distinct camera perspectives. Artificial intelligence combined with the IoT can provide an intelligent system as illustrated in FIGURE 2, which can be utilized in different surveillance applications.



**FIGURE 2** Artificial Intelligence combined with Internet of Things for Human Detection and other applications.

In this work, an automated human detection system is introduced for a complex industrial environment. The intelligent surveillance system utilized deep learning detection architectures i.e., (Faster Region Convolutional Neural Network) Faster R-CNN [4], (Single Shot MultiBox Detector) SSD [5], (You Only Look Once) YOLOv3 [6], as a baseline for top view human detection. All deep learning models are firstly tested on SCOVIS data set [7], [8], containing images of people captured in a complex industrial environment. As these architectures are previously trained on frontal view data sets, and the person's visual features from the top view are distinctive. Therefore, all three architectures are additionally trained on the SCOVIS data set, and by leveraging transfer learning, the newly formed layer is interfaced with pre-trained architectures. The additional training improves the accuracy results of the top view human detection system in cluttered and complex industrial environments. This effort might be considered the first approach in which different deep learning-based architectures are trained and tested utilizing a complex top view industrial data set.

Generally, the primary purpose of the work is given as;

- To present an automated IoT-based human detection system for a complex industrial environment. The developed system utilizes deep learning architectures along with the Internet of Things for top view human detection. .
- To study the generalization performance of deep learning architectures by testing on an entirely different data set, i.e., top view industrial data set.
- To further boost the performance, the detection architectures are trained for the SCOVIS data set; by applying deep transfer learning, the trained layer is interfaced with the pre-trained architectures.

- To compare the detection results of trained and pre-trained architectures.
- To explore the importance of top view human detection in the industrial environment with possible future direction. .

The work illustrated in the article is arranged as follows: a concise review of several top view human detection techniques is addressed in SECTION 2. The deep learning architectures explored for top view human detection along with transfer learning are elaborated in SECTION 3. The detail of the data set utilized for testing and training is explained in SECTION 4. SECTION 5 provides a comprehensive discussion of experimentation and results. Lastly, the conclusion of the work is discussed in SECTION 6 with possible future trends.

## 2 | RELATED WORK

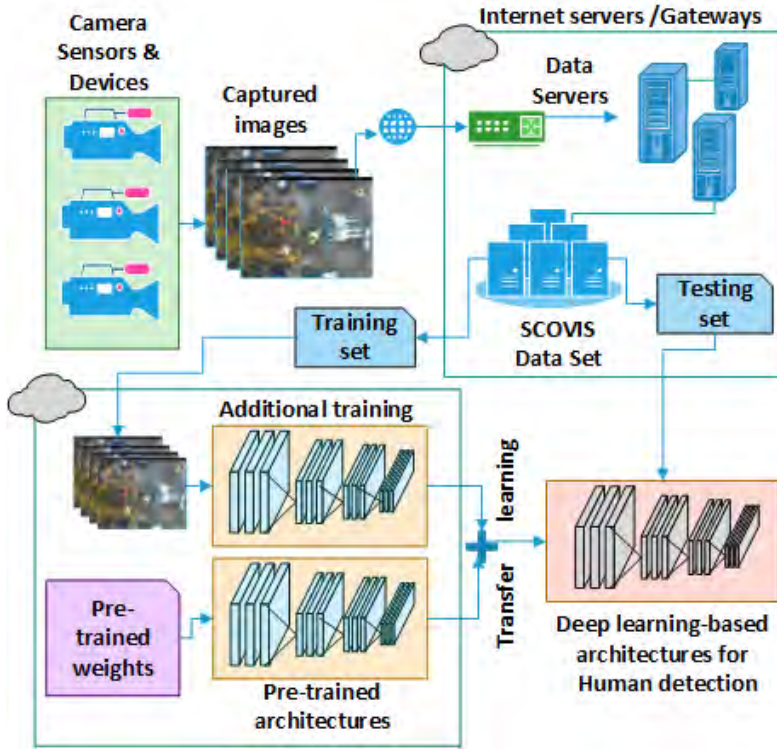
In literature, mostly researchers have studied the perspective of the front view camera for human detection while some have performed human detection using top view, by considering head and shoulder information, or rarely complete information of the human body. Mostly human detection techniques proposed for top view focused on different background subtraction and segmentation approaches, e.g., Iguernaissi et al. [9]. These techniques detect humans using head region information in different constrained environments, e.g., [10]. Few of them used feature-based methods for top view human detection, such as in [11] proposed an efficient approach using HoG (Histogram of Oriented Gradient) characteristics. Ahmed et al. [12] [8] introduced a Rotated-HoG algorithm utilizing industrial images (SCOVIS data set) [7] for human detection. [2], utilized a top view wide-angle camera and developed an algorithm that detects humans from different angles using variable-sized bounding boxes. In [13], authors presented a feature-based approach for people detection from a top perspective in an industrial environment. [12] utilized feature information for person detection and tracking, using the SCOVIS data set. Ullah et al. [14] also used blob based approach and presented a rotation-invariant human tracking and detection approach for top view monitoring. The scholars in [1] presented an efficient method for top view person detection named a rotated HoG based approach using an SVM classifier.

Some literature studies are also utilizing deep learning-based techniques for human detection via applying the fisheye camera [15]. Ahmed et al., [16] applied two deep learning techniques to detect multi-class objects from a top perspective. Ahmad et al., [17] applied the convolutional neural network-based method for top view human detection and tracking. A comparison of different segmentation techniques applied for top view human detection is presented by [18]. [19] presented a collaborative robotics-based top view surveillance system for multiple object tracking and detection. Ahmed et al., also made another effort for top view multiple people detection and tracking utilizing YOLOv3 and Deep SORT technique. It is concluded that researchers used various top view features of the human body for detection purposes, such as shape, color, and size. Though, most of these developed techniques are based on traditional feature-based techniques.

## 3 | METHODOLOGY

This work has leveraged artificial intelligence and introduced a system for automated human detection for a complex industrial environment. The system utilizes object detection techniques based on deep learning architectures for human detection. The overall system is presented in Figure 3. We firstly practiced the pre-trained deep learning architectures i.e., Faster R-CNN [4], SSD [5], and YOLOv3 [6], for the detection of people in an industrial data set.

As architectures are previously trained on the MS-COCO data set [20] in which mostly the person/human body is considered from a frontal and side view perspective. Therefore, to increase the model's accuracy for top view human detection, training is performed using the SCOVIS data set. The transfer learning significantly improves the detection results for the SCOVIS data set. The results of both trained and pre-trained architectures are evaluated and further processed for monitoring and management purposes.

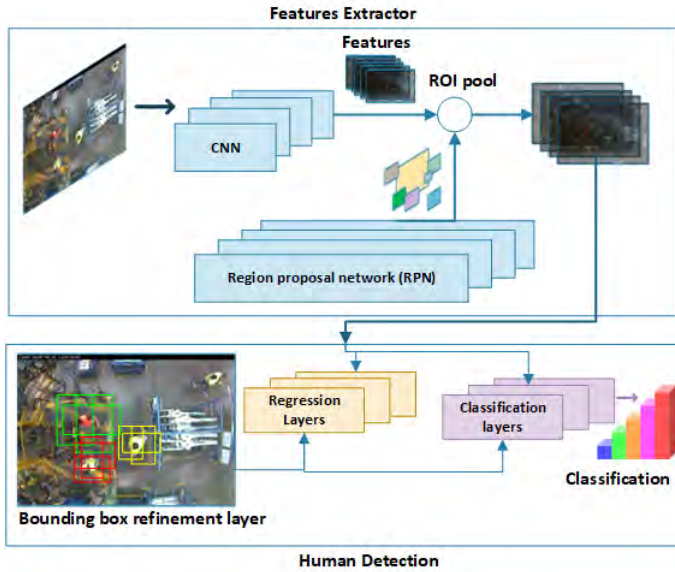


**FIGURE 3** An IoT-based human detection system for a complex industrial environment. Different deep learning architectures have been utilized for top view human detection. Pre-trained architectures are additionally trained for the top view.

### 3.1 | Top view Human Detection using Faster-RCNN:

In this work, firstly, we implemented Faster R-CNN [4] for top view human detection as presented in FIGURE 4. It outputs the rectangular bounding box containing (information about coordinates, height, and width of the bounding box) along with the confidence score value and class label. It is also known as two-stage detectors; the first stage generates region anchors through Region Proposal Network (RPN). The next stage is utilized for object classification (e.g., human) that uses detected anchor regions and extracts bounding box information. To further improve detection results, we additionally trained the architecture for the SCOVIS data set. The overall process shown in FIGURE 4 mainly has three steps:

- The features for the input image are extracted through convolution layers. The extracted features are further utilized for the generation of feature maps.
- Using the sliding window approach, anchor or region boxes are generated. To specify an object/human's presence in the image, these anchor boxes are further refined.
- In the final step, applying a small network, anchors/detected bounding boxes are refined, and the loss function is computed that decides the most suitable anchors regions.



**FIGURE 4** The general architecture of Faster-RCNN, used for human detection in top view industrial environment.

In FIGURE 4, it can be observed that top view images are passed through the network of conventional layers. We used Inceptionv3 based model backbone for the extraction of image features. The detected bounding box coordinates values are mathematically provided as[21]:

$$\begin{aligned}
 t_x &= \frac{(x-x_a)}{w_a}, & t_y &= \frac{(y-y_a)}{h_a} \\
 t_w &= \log \frac{w}{w_a}, & t_h &= \log \frac{h}{h_a} \\
 t_x^* &= \frac{(x^*-x_a)}{w_a}, & t_y^* &= \frac{(y^*-y_a)}{h_a} \\
 t_w^* &= \log \frac{w^*}{w_a}, & t_h^* &= \log \frac{h^*}{h_a}
 \end{aligned} \tag{1}$$

In Equation 1, central coordinates values of bounding box are represented with  $x$ ,  $y$ , height and width is denoted with  $h$  &  $w$ , while the predicted bounding box is denoted by  $x$ ,  $x_a$  &  $x^*$ , respectively. The ( $IOU$ ) Intersection Over Union

approach is employed to estimate how ground-truth  $GT$  bounding boxes overlap with anchors regions. Furthermore, a threshold value is determined for a region, including an object being human or other objects/backgrounds. The probability for human class based upon  $IOU$  is provided in the equation below:

$$IOU = \frac{B_{\text{ounding Box}}(\text{anchor}) \cap GT}{B_{\text{ounding Box}}(\text{anchor}) \cup GT} \begin{cases} > 0.7 = \text{person} \\ < 0.3 = \text{other objects} \end{cases} \quad (2)$$

After selecting anchor regions, the loss function is applied for fine-tuning of detected anchors at the end of RPN. The regression and classification loss function is mathematically defined as [4]:

$$\mathcal{L}(\{p_i\}, \{t_i\}) = \frac{1}{N_{cls}} \sum_i \mathcal{L}_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* \mathcal{L}_{reg}(t_i, t_i^*) \quad (3)$$

In Equation 3,  $i$  represents the anchor regions index,  $p_i$  is the predicted probability of the human or person class and expressed as  $\lambda$ . The  $p_i^*$  indicates ground truth; if its value is equal to 1, formerly anchor region relates to the positive class means an object is detected; else, if equal to 0, the anchor relates to a negative class or also referred to as other objects.  $t_i$  represents a vector applied to modify the predicted bounding box's coordinates, whereas  $t_i^*$  is employed to describe the ground truth-bounding box's coordinates. RPN is previously trained to provide (Regions of Interest) RoIs over convolution feature maps. Once from the ROI pool layers, the similar size feature maps are extracted. The loss function is utilized for classification and regression. At the output, the detected bounding box is produced containing humans with a class score value.

### 3.2 | Top view Human Detection using SSD:

The second deep learning architecture utilized for top view human detection is SSD-MobileNetV2 [5]. The overall architecture is provided in FIGURE 5. The input images are fed into the detection module, which is based on SSD-MobileNetV2. It consists of a  $3 \times 3$  depth-wise convolution layers followed by a  $1 \times 1$  point-wise convolution layers. The SSD model firstly extracts object features and then used convolution filters to detect the object. To perform classification and localization regression, different feature maps are used, and a set of default boxes are specified to each cell of the feature maps, as depicted in FIGURE 6.

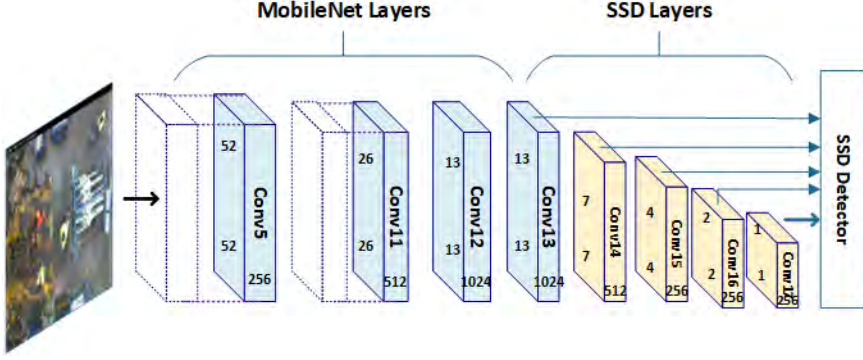
For every feature map, the model provides four different-sized bounding boxes. As shown in FIGURE 6b and FIGURE 6c,  $m \times n$  size feature map is obtained with  $p$  channels. Through convolution layers. For each detected location at feature map,  $k$  bounding boxes are extracted, having different aspect ratios and sizes, but the aspect ratio of the default bounding box is the same expressed as [5]:

$$s_k = s_{min} + \frac{s_{max} - s_{min}}{m - 1} (k - 1), \quad k \in [1, m] \quad (4)$$

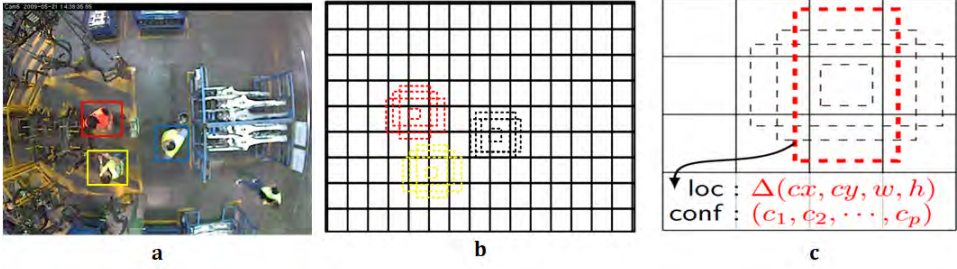
In Equation 4 for prediction of  $m$  feature maps, scale value  $s_k$  is estimated for the  $k^{th}$  feature map. The  $s_{min}$  value is equal to 0.2, and  $s_{max}$  is equal to 0.9 [5], which determines that the highest and lowest layer are scaled at 0.9 and 0.2 respectively. For each  $s_k$ , five different aspect ratios  $a_r$  (non-square root bounding boxes) are considered, provided as;



$$a_r \in \{1, 2, 3, 4, \frac{1}{2}, \frac{1}{3}\} \quad (w_k^a = s_k \sqrt{a_r}) \quad (h_k^a = s_k \sqrt{a_r}) \quad (5)$$



**FIGURE 5** The general architecture of top view human detection in industrial scene using SSD-MobileNetv2.



**FIGURE 6** Multiple extracted bounding boxes detected for the target object (human). (a) Input image with ground truth boxes (training). Feature maps with various scale sizes are shown in b and c. (based on [5]).

$w$  and  $h$  in Equation 5 are the width and height of each detected bounding box. For aspect ratio 1:1, one default bounding  $s'_k = \sqrt{s_k, s_{k+1}}$  is added. Thus, the resulting six bounding boxes are estimated from the above equations with different aspect ratios. The loss function is computed as [5]:

$$\mathcal{L}(x, c, l, g) = \frac{1}{N} (\mathcal{L}_{conf} + \alpha \mathcal{L}_{loc}) \quad (6)$$

In Equation 6  $N$  represents a number of bounding boxes matched, and  $\alpha$  describes the stabilized weights among two loss functions. The  $\mathcal{L}_{conf}$  expresses the confidence loss function and  $\mathcal{L}_{loc}$  represents localization loss function. The  $\mathcal{L}_{conf}$  is given as [5]:

$$\mathcal{L}_{conf}(x, c) = - \sum_{i \in pos} x_{ij}^p \log c_i^p - \sum_{i \in neg} \log c_i^o \quad (7)$$



The value of  $c_i^p$  is computed as;

$$c_i^p = \frac{\exp(c_i^p)}{\sum_p \exp(c_i^p)} \quad (8)$$

The confidence loss function (Softmax) [22] is estimated as the loss across different class confidence values  $c$ . For matching of the  $j$ th ground truth box of class  $p$  with predicted  $i$ th box, the value of  $x_{ij}^p = 1, 0$  is used. The  $\mathcal{L}_{loc}$  shown in Equation 6 is determined as [5]:

$$\begin{aligned} \mathcal{L}_{loc}(x, l, g) &= \sum_{ipos} \sum_{me \in \{c_x, c_y, w, h\}} x_{ij}^k L_1^{smooth}(l_i^m - g_j^m) \\ g_j^{cx} &= (g_j^{cx} - d_i^{cx} / d_i^w) g_j^{cy} = (g_j^{cy} - d_i^{cy} / d_i^h) \\ g_j^w &= \log \frac{g_j^w}{d_i^w} g_j^h = \log \frac{g_j^h}{d_i^h} \end{aligned} \quad (9)$$

Equation 9 provides the localization loss measured between the ground truth box  $g$  and predicted bounding box  $l$ . Where  $x_{ij}^k$  signifies the  $i$ th matched bounding box coordinates with  $j$ th ground truth coordinates of target object. The  $h, w$  are the bounding box's height and width, while  $c_x, c_y$  are the center points.

### 3.3 | Top view Human Detection using YOLOV3:

In this work, YOLOv3 is also additionally trained for the SCOVIS data set. The new trained layer is further embedded with the previously learned architecture as demonstrated in FIGURE 3, two weights are joined, and a new detection system/architecture is produced that significantly enhances human detection results for industrial data set. The YOLOv3 adopted a one-stage architecture to predict identical bounding boxes and class probability for the entire image. The convolution layers are utilized for the image's feature extraction, whereas to determine the class probability and prediction, fully connected layers are applied. The general architecture is visualized in FIGURE 7 (adopted from literature [23]), the image is segmented into regions of  $S \times S$  for human detection, also named grid cells, those are correlated with class probabilities and bounding box predictions. Each grid cell predicts the possibility; either the center of the detected human lies within the cell or not. The confidence values and bounding boxes are predicted for all positive detections if the prediction is positive. The overall process of human detection using YOLOv3 architecture is visualized in FIGURE 8. It implies the strength of the bounding box detected as a human or person and determined as:

$$C_{conf\ person} = Pr_{person} \times IOU(P_{red}, T_{ruth}) \quad (10)$$

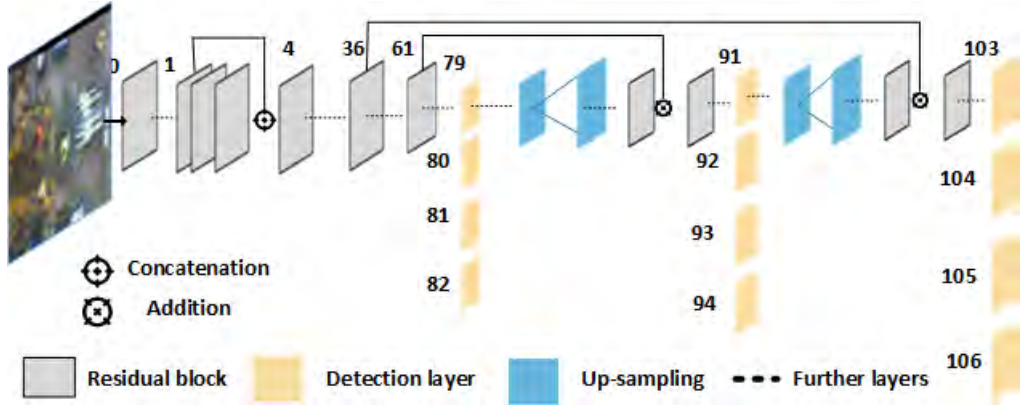
In Equation 10,  $Pr_{person}$  represents either human exists in predicted bounding box or not (1 is for yes & 0 is for not). For the intersection of the actual and predicted bounding box,  $IOU(P_{red}, T_{ruth})$  is utilized. It is mathematically provided as:

$$IOU(P_{red}, T_{ruth}) = \frac{area_{B_{ox}T \cap B_{ox}P}}{B_{ox}T \cup B_{ox}P} \quad (11)$$

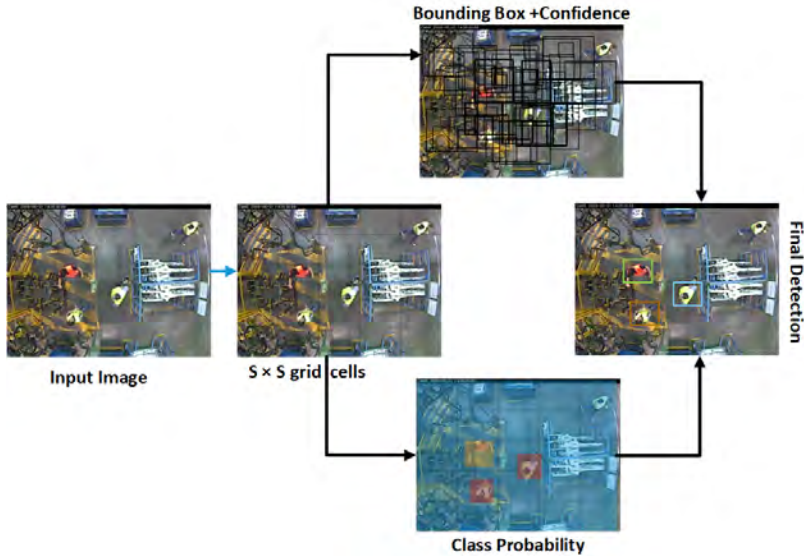
For top view human detection, the appropriate regions are chosen and predicted. The confidence value is utilized in

order to attain the desired bounding box after prediction. Ultimately, the loss function is determined for the detected human bounding box, a combination of regression and classification loss. Nevertheless, only one object is considered in this work, i.e., human/person. Accordingly, the loss function is mathematically given as:

$$loss(person) = \mathcal{L}_c + \mathcal{L}_{IOU} \quad (12)$$



**FIGURE 7** The architecture of YOLOv3 used for human detection in top view industrial scene.



**FIGURE 8** Human detection using YOLOv3.

In Equation 12,  $\mathcal{L}_c$  denotes loss of the predicted and  $\mathcal{L}_{IOU}$  is utilized for estimating a loss of actual bounding box

coordinates. The loss function of coordinates  $\mathcal{L}_c$  is mathematically presented as [23]:

$$\mathcal{L}_c = \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^B I_{ij}^{person} [(x_i - x_i^*)^2 + (y_i - y_i^*)^2 + (\sqrt{w_i} - \sqrt{w_i^*})^2 + (\sqrt{h_i} - \sqrt{h_i^*})^2] \quad (13)$$

In Equation 13  $\lambda_{coord}$  is utilized as scale parameters for bounding box coordinates predictions, the value of  $\lambda_{coord} = 5$ . For detected bounding box, predicted positions  $h_i, w_i, x_i, y_i$  in  $i_{th}$  cell is determined. The bounding box's actual positions in the  $i_{th}$  cell is represented as  $h_i^*, w_i^*, x_i^*, y_i^*$ . The  $\mathcal{L}_{IOU}$  of  $IOU_{pred}^{truth}$  is determined as:

$$\begin{aligned} \mathcal{L}_{IOU} &= \sum_{i=0}^{S^2} \sum_{j=0}^B I_{ij}^{person} [(\xi_i - \xi_i^*)^2 + \\ \lambda_{no-person} &\sum_{i=0}^{S^2} \sum_{j=0}^B I_{ij}^{no-person} [(\xi_i - \xi_i^*)^2 \end{aligned} \quad (14)$$

In Equation  $\lambda_{no-person}$  is used to present classification error; furthermore, the value of the confidence in  $i_{th}$  grid cell is represented as  $\xi_i$ , and  $\xi_i^*$  in the predicted and actual sliding window.  $I_{person}$  describes either the person is identified in  $j_{th}$  bounding box of  $i_{th}$  grid cell or not. If human is present in  $j_{th}$  bounding box &  $i_{th}$  grid cell, thus the function value is equivalent to 1; else, it is equal to 0.

## 4 | TRAINING AND TESTING

To train and test the model, the SCOVIS (Self-Control Cognitive Video Supervision) data set is used [7]. This data is recorded in a real-world industrial environment (NISSAN Motor Iberica SA (NMISA), in Barcelona, Spain). The images are captured from the top view perspective in uncontrolled circumstances, in which lightning conditions are varying. The working field in the environment is limited, but humans are freely moving without any restrictions; thus, there is variation in the person's visual features with regard to size, scale, orientation, and pose. The images were captured with a resolution of  $640 \times 480$  pixels, utilizing a camera positioned at five meters height. The recording of images has been made with high gain at a low luminance level, and JPEG compression has almost 50 KB size each. A total of 4000 sample images have been used in this work, of which 1000 are utilized for training and 3000 for testing. For training, the model's learning rate is initialized at  $2.5e4$  and decreased with factor 10 at the end of the 90 and 120 epoch. The same training parameters include random scaling, random flip (within 0.6 to 1.3), and Adam [24] are utilized to optimize the overall objective.

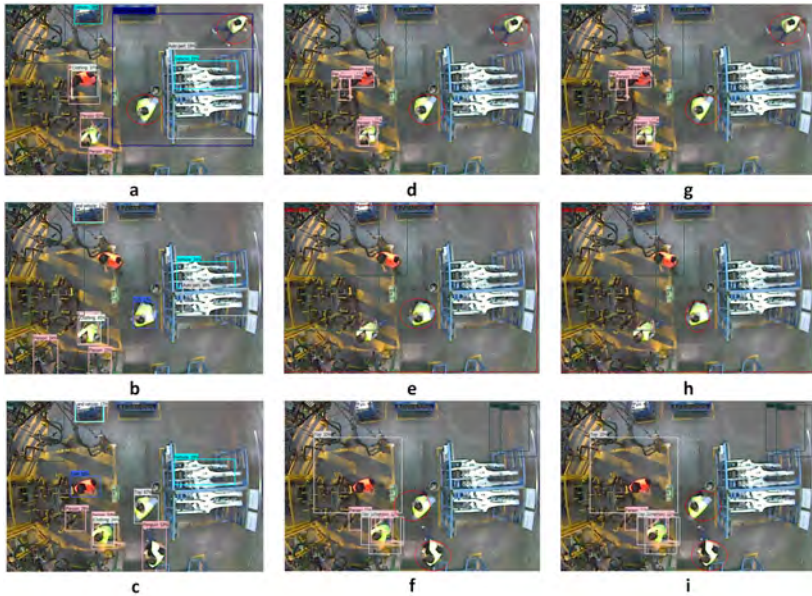
## 5 | EXPERIMENTAL RESULTS

The detail of the experimental results has been discussed in this section. The section is categorized into two subsections: detection results and performance evaluation results. The first section explains human detection results in an industrial environment using pre-trained architectures and newly trained architectures. The performance evaluation and comparison of the architectures with other detection architectures used for human detection is presented in the second section. The human detection results are further classified into two subsections; firstly, the pre-trained

architectures are tested using SCOVIS test images. As the data set is recorded from a top view perspective in a complex industrial environment, the result's detection architectures are not as good as expected. Therefore, the model is further trained using the SCOVIS data set.

### 5.1 | Detection results of the pre-trained architectures tested on SCOVIS data set.

The pre-trained architectures results after testing for human detection using the top view data set (SCOVIS) in an industrial environment can be examined in FIGURE 9. From the sample test images, it is observed that the scene is quite complex in comparison with the data set used by the pre-trained architectures. Also, as a top view perspective is considered, the human's visual features are varying throughout the scene. In comparison with [16], the pre-trained architectures could not give good results because of complex background conditions. In FIGURE 9, the first column presents the detection results of Faster-RCNN, the second explains SSD, and the third shows the results of the YOLOv3 detection model. In FIGURE 9a, it can be viewed that three persons are detected (shown in pink bounding box), while the other two, one exactly below the camera and two at the bottom right, are not detected (marked with red circles).



**FIGURE 9** Detection results of pre-trained architectures after testing on SCOVIS data set. The first column shows Faster-RCNN, the second column shows SSD, and the third shows the YOLOv3 detection model results. The red circles show the not detected results, while the detected results can be visualized with pink rectangular bounding boxes for humans.

In FIGURE 9b & FIGURE 9c, people working below the camera are not detected by the pre-trained architectures marked with red circles in the sample image. In the case of SSD and YOLOv3, the detection results are also not good; for example, in the case of SSD as shown in FIGURE 9d, FIGURE 9e, and FIGURE 9f, there are some cases in which human is detected, but in most of the cases, the architecture gives not detected results. Also, for the YOLOv3

detection model, the results are the same as SSD, depicted in FIGURE 9g, FIGURE 9h, & FIGURE 9i. There are also some miss detection results, in which the human body is detected as other objects, e.g., in FIGURE 9c human body is detected as fish, while in FIGURE 9f & FIGURE 9i, the human body is categorized as a toy. The pre-trained architecture in FIGURE 9 also detects machinery as the human body, as shown at the bottom of the sample image. In mostly cases, the architectures detected the person as another object or give multiple detections for a single human body. In sample images, mostly the appearance of people working below the camera is similar; that is why the pre-trained architecture does not produce good results. The individual is detected from the top view, represented in the pink bounding box with class label and score. Most of the detection might occur because the human's visual appearance is somehow the same as the frontal view.

## 5.2 | Detection results of deep learning architectures additionally trained and tested on SCOVIS data set.

After training all three architectures on the SCOVIS data set, testing is performed using the same sample images. Detection accuracy of all three detection models is enhanced, as discussed in SECTION 4. Applying transfer learning, architectures are trained using 1000 images of humans collected in a cluttered industrial environment. For training, the batch size of 128 and the epochs number = 100 is employed. The training accuracy and loss are achieved at the end of the 90<sup>th</sup> epoch. The accuracy and loss of deep learning architectures are described in FIGURE 10(a) and FIGURE 10(b), respectively. In FIGURE 10(a), the loss convergence easily determines the effectiveness of deep learning architectures. The value of error is near to 0.2 or 0.3 after 20<sup>th</sup> epoch. Similarly, in FIGURE 10(b), the training accuracy reveals that the YOLO and SSD perform better in comparison with Faster-RCNN. It can also be noted that the model's training accuracy is almost 95% after the 20<sup>th</sup> epoch.



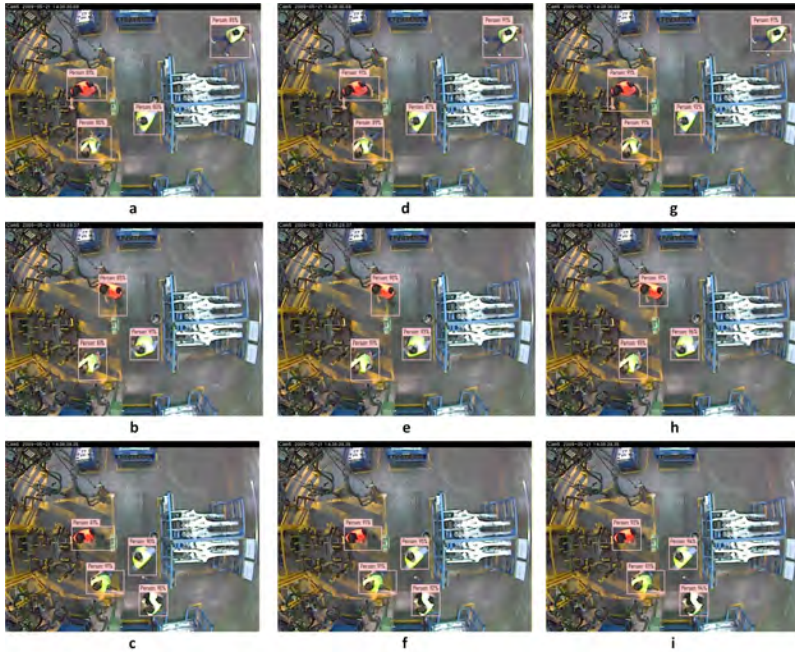
**FIGURE 10** (a) Training Loss of Faster-RCNN, SSD and YOLOv3. (b) Training Accuracy of Faster-RCNN, SSD and YOLOv3.

After training, architectures are again tested for human detection using top-view images, as visualized in FIGURE 11. The results demonstrate that additional training significantly improved the detection results for the top view industrial data set. The architectures are tested using the same images used in the previous section. In FIGURE 11, it can be observed that deep learning architectures effectively detect multiple people in the industrial scene. Since there are many people in different places and their visual characteristics are also different from one another, all three architectures detect bounding boxes of varying sizes (width and height) for each target human; for example, in the case of



Faster-RCNN, three people are adequately detected as shown in FIGURE 11a, FIGURE 11b, & FIGURE 11c.

In the same way, for SSD, in FIGURE 11d, FIGURE 11e, & FIGURE 11f, three people exactly below the camera are detected effectively, which are not detected by the pre-trained architecture, as displayed in FIGURE 9d, FIGURE 9e & FIGURE 9f. In FIGURE 11c, four people with varying body orientation working specifically below the camera is accurately detected and classified by the deep learning model with a class score of more than 90%. The model detects the same in the case of FIGURE 11d, two people with the same color clothes but different visual appearances. Almost the same results have been obtained for YOLOv3 as shown in FIGURE 11g, FIGURE 11h, & FIGURE 11i. It can be concluded from FIGURE 11 that the human body's visual characteristics in all sample images are not alike. The architectures are trained on the SCOVIS data set obtained from the top view and interfaced with the pre-trained architectures. It enhances human detection accuracy; as seen in the sample test images, people's movement over the scene and their location concerning camera positions are varied; appearance size and scale alter. Despite this, the architectures with transfer learning detect people without failure and modify the bounding boxes' scale and size according to their body scale and sizes.



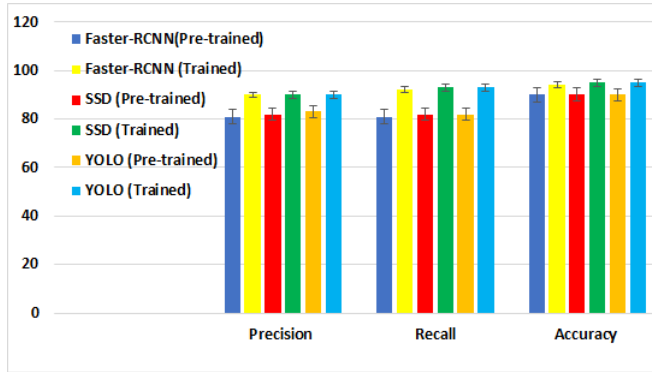
**FIGURE 11** Detection results after training architectures on the SCOVIS data set. The first column shows Faster-RCNN, the second column shows SSD, and the third shows the YOLOv3 detection model results. The red circles show the not detected results, while the detected results can be visualized with pink rectangular bounding boxes for humans.

### 5.3 | Performance Evaluation

Deep learning architectures detection results are evaluated using various evaluation parameters, such as  $tp$ ,  $fp$ ,  $fn$ , &  $tn$ . The Precision, Recall, and Accuracy is calculated for deep learning architectures. In FIGURE 12, the Precision, Recall,

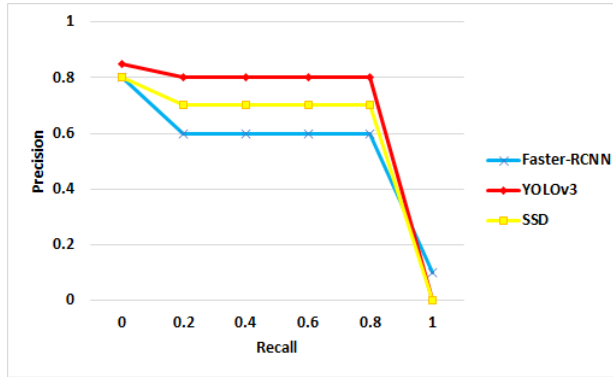


and Accuracy of the pre-trained and trained architectures utilizing the SCOVIS data set can be examined. The standard deviation and mean are used to estimate the average Precision, Recall, and Accuracy.



**FIGURE 12** Precision, Recall, and Accuracy of deep learning architectures.

The Precision-Recall (PR) curve shows the performance measure between positive detection and positive predictions. Using the Precision value on the y-axis and Recall value on the x-axis, the curve is plotted as depicted in FIGURE 13. It can be examined after training deep learning architectures; the performance is improved. Among all, the YOLOv3 and SSD trained on the SCOVIS data set performs better in contrast with Faster-RCNN.

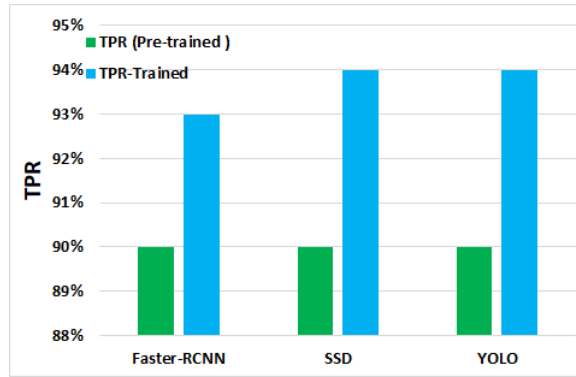


**FIGURE 13** Recall with respect to the Precision (PR curves) of deep learning architectures.

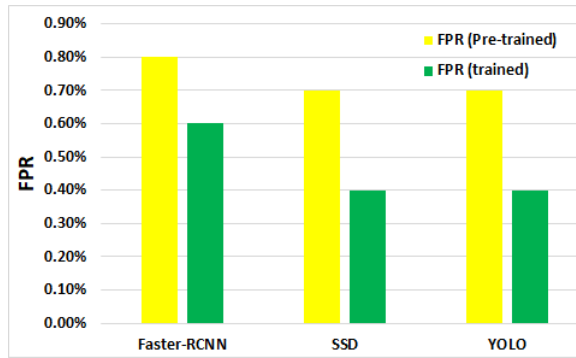
In FIGURE 14, True Positive Rate (TPR) and False Positive Rate (FPR) is utilized to determine the results. It can be observed that almost all deep learning architectures provided good results both in terms of low false-positive and high detection rates. The detection performance of YOLOv3 and SSD is slightly better than Faster-RCNN. For top view human detection in an industrial setup, the Faster-RCNN achieves TPR of 84% and 94% without and with transfer learning. While the TPR of SSD and YOLO is 86% and 94% without and with transfer learning, respectively.

We also plotted FPR without and with transfer learning is also shown in FIGURE 15. The FPR of Faster-RCNN is 0.80% and 0.60%. In the case of SSD and YOLO, the FPR of 0.70% and 0.4% without and with transfer learning.

From comparison results, it can be noticed that the accuracy of the detection architectures is increased after training



**FIGURE 14** True Positive Rate (TPR) of deep learning architectures for top view human detection.



**FIGURE 15** False Positive Rate (FPR) of deep learning architectures for top view human detection.

and applying transfer learning. Researchers in literature used Average Precision,  $AP$  value, to investigate architectures performance; therefore, to consistent with researchers, we also calculated the  $AP$ , to examine the performance of architectures. In the T ABLE10, we show the comparison of the detection architectures with each other.

**TABLE 1** Comparison results of deep learning architectures using Average Precision  $AP$  value.

S.No	Detection Model	Average Precision $AP$
1	Faster-RCNN (pre-trained)[4]	84 %
2	Faster-RCNN (trained)[4]	90 %
3	SSD (pre-trained) [5]	86 %
4	SSD (trained) [5]	93 %
5	YOLOv3 (pre-trained) [23]	86 %
6	YOLOv3 (trained) [23]	90 %

## 6 | CONCLUSION AND FUTURE DIRECTIONS

This article featured a human detection system based on artificial intelligence for a complex industrial environment. Different deep learning architectures are explored for human detection, i.e., Faster-RCNN, SSD, and YOLOv3. Depth learning architectures have detected bounding boxes of varying sizes with class labels and scores in the top view's images. Firstly, the pre-trained architectures are practiced for testing, as the appearance of the human body in the top perspective is varying. Therefore, we trained architectures on the SCOVIS data set (a top view industrial data set) in which neither any supposition about the person's visibility and pose nor any constraint on the environment is made. We interfaced with the new layer formed with the previously trained architectures re-tested it for the SCOVIS data set through the deep transfer learning application. Additional training enhances the model's detection results. The model delivers detection accuracy of 90% and 95% with a pre-trained and trained model, respectively. The True Positive Rate (TPR) is 93%, for Faster-RCNN, and 95% for SSD, and YOLOv3, respectively. At the same time, False Positive Rate (FPR) is 0.60% and 0.4% for Faster-RCNN, SSD, and YOLO, respectively. In the future, this work may be expanded to other deep learning paradigms. We could also use different tracking algorithms to monitor people in the industrial scene from the top view perspective. This data set may also help to monitor the different activities of those peoples working in the industry.

## references

- [1] Ahmed I, Ahmad M, Nawaz M, Haseeb K, Khan S, Jeon G. Efficient topview person detector using point based transformation and lookup table. *Computer Communications* 2019;147:188–197.
- [2] Ahmed I, Adnan A. A robust algorithm for detecting people in overhead views. *Cluster Computing* 2018 Mar;21(1):633–654. <https://doi.org/10.1007/s10586-017-0968-3>.
- [3] Ahmad M, Ahmed I, Ullah K, Khan I, Khattak A, Adnan A. Energy Efficient Camera Solution for Video Surveillance. *International Journal of Advanced Computer Science and Applications* 2019;10(3). <http://dx.doi.org/10.14569/IJACSA.2019.0100367>.
- [4] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R, editors. *Advances in Neural Information Processing Systems* 28 Curran Associates, Inc.; 2015.p. 91–99.
- [5] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. Ssd: Single shot multibox detector. In: *European conference on computer vision* Springer; 2016. p. 21–37.
- [6] Redmon J, Farhadi A. Yolov3: An incremental improvement. *arXiv preprint arXiv:180402767* 2018;.
- [7] Doulamis A, Kosmopoulos D, Sardis M, Varvarigou T. An architecture for a self configurable video supervision. In: *Proceedings of the 1st ACM workshop on Analysis and retrieval of events/actions and workflows in video streams*; 2008. p. 97–104.
- [8] Ahmed I, Carter JN. A robust person detector for overhead views. In: *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*; 2012. p. 1483–1486.
- [9] Iguernaissi R, Merad D, Drap P. People Counting based on Kinect Depth Data. In: *Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods - Volume 1: ICPRAM, INSTICC, SciTePress*; 2018. p. 364–370.
- [10] Tseng T, Liu A, Hsiao P, Huang C, Fu L. Real-time people detection and tracking for indoor surveillance using multiple top-view depth cameras. In: *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*; 2014. p. 4077–4082.

- [11] Pang Y, Yuan Y, Li X, Pan J. Efficient HOG human detection. *Signal Processing* 2011;91(4):773–781.
- [12] Ahmed I, Ahmad A, Piccialli F, Sangaiah AK, Jeon G. A robust features-based person tracker for overhead views in industrial environment. *IEEE Internet of Things Journal* 2018;5(3):1598–1605.
- [13] Ahmed I, Ahmad M, Adnan A, Ahmad A, Khan M. Person detector for different overhead views using machine learning. *International Journal of Machine Learning and Cybernetics* 2019 Oct;10(10):2657–2668. <https://doi.org/10.1007/s13042-019-00950-5>.
- [14] Ullah K, Ahmed I, Ahmad M, Rahman AU, Nawaz M, Adnan A. Rotation invariant person tracker using top view. *Journal of Ambient Intelligence and Humanized Computing* 2019;p. 1–17.
- [15] Ertler C, Possegger H, Opitz M, Bischof H. Pedestrian Detection in RGB-D Images from an Elevated Viewpoint. In: Kropatsch W, Janusch I, Artner N, editors. *Proceedings of the 22nd Computer Vision Winter Workshop Austria: TU Wien, Pattern Recognition and Image Processing Group*; 2017. .
- [16] Ahmed I, Din S, Jeon G, Piccialli F. Exploring Deep Learning Models for Overhead View Multiple Object Detection. *IEEE Internet of Things Journal* 2020;7(7):5737–5744.
- [17] Ahmad M, Ahmed I, Khan FA, Qayum F, Aljuaid H. Convolutional neural network-based person tracking using overhead views. *International Journal of Distributed Sensor Networks* 2020;16(6):1550147720934738.
- [18] Ahmed I, Ahmad M, Khan FA, Asif M. Comparison of Deep-Learning-Based Segmentation Models: Using Top View Person Images. *IEEE Access* 2020;8:136361–136373.
- [19] Ahmed I, Din S, Jeon G, Piccialli F, Fortino G. Towards collaborative robotics in top view surveillance: A framework for multiple object tracking by detection using deep learning. *IEEE/CAA Journal of Automatica Sinica* 2020;.
- [20] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: Common Objects in Context. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, editors. *Computer Vision – ECCV 2014 Cham: Springer International Publishing*; 2014. p. 740–755.
- [21] Girshick R, Donahue J, Darrell T, Malik J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; 2014. p. 580–587.
- [22] Erhan D, Szegedy C, Toshev A, Anguelov D. Scalable object detection using deep neural networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*; 2014. p. 2147–2154.
- [23] Redmon J, Divvala S, Girshick R, Farhadi A, You Only Look Once: Unified, Real-Time Object Detection; 2016.
- [24] Kingma DP, Ba J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* 2014;.