International Journal of Machine Learning and Cybernetics (2021) 12:3053–3067
https://doi.org/10.1007/s13042-020-01220-5

ORIGINAL ARTICLE

Top view multiple people tracking by detection using deep SORT and YOLOv3 with transfer learning: within 5G infrastructure

Imran Ahmed[1] · Misbah Ahmad[1] · Awais Ahmad[2] · Gwanggil Jeon[3,4]

Received: 26 June 2020 / Accepted: 1 October 2020 / Published online: 27 October 2020

Abstract

Nowadays, 5G profoundly impacts video surveillance and monitoring services by processing video streams at high-speed with high-reliability, high bandwidth, and secure network connectivity. It also enhances artificial intelligence, machine learning, and deep learning techniques, which require intense processing to deliver near-real-time solutions. In video surveillance, person tracking is a crucial task due to the deformable nature of the human body, various environmental components such as occlusion, illumination, and background conditions, specifically, from a top view perspective where the person's visual appearance is significantly different from a frontal or side view. In this work, multiple people tracking framework is presented, which uses 5G infrastructure. A top view perspective is used, which offers broad coverage of the scene or field of view. To perform a person tracking deep learning-based tracking by detection framework is proposed, which includes detection by YOLOv3 and tracking by Deep SORT algorithm. Although the model is pre-trained using the frontal view images, even then, it gives good detection results. In order to further enhance the accuracy of the detection model, the transfer learning approach is adopted. In this way, a detection model takes advantage of a pre-trained model appended with an additional trained layer using top view data set. To evaluate the performance, experiments are carried out on different top view video sequences. Experimental results reveal that transfer learning improves the overall performance, detection accuracy, and reduces false positives. The deep learning detection model YOLOv3 achieves detection accuracy of 92% with a pre-trained model without transfer learning and 95% with transfer learning. The tracking algorithm Deep SORT also achieves excellent results with a tracking accuracy of 96%.

Keywords 5G · Deep learning · YOLOv3 · Deep SORT · Transfer learning · Person detection and tracking · Top view

1 Introduction

Traditional video surveillance deployments were reliant on complex fiber and cable connections, which are expensive to implement and manage at scale. Nowadays, wireless connectivity is utilized for video transmissions in video surveillance analysis. 5G networks improve mobile broadband capabilities and provide high bandwidth requirements for high-resolution video cameras needed for delivering the quality of services [1]. It also enhances video security surveillance systems and applications, particularly in person tracking and detection. It has a key capability for many video surveillance applications such as crowd analysis [2, 3], robotics [4], security analysis[5, 6], autonomous or self-driving vehicles [7, 8], Human-computer interaction (HCI) [9, 10], face recognition [11] location and navigation and most importantly person tracking and detection. However, it is likewise considered one of the challenging tasks for researchers because

✉ Gwanggil Jeon
ggjeon@gmail.com

Imran Ahmed
imran.ahmed@imsciences.edu.pk

Misbah Ahmad
misbahahmad4872@gmail.com

Awais Ahmad
aahmad.marwat@gmail.com

[1] Center of Excellence in Information Technology, Institute of Management Sciences, Hayatabad, Peshawar, Pakistan

[2] Department of Computer Science, Air University, Islamabad 44000, Pakistan

[3] School of Electronic Engineering, Xidian University, Xi'an 710071, China

[4] Department of Embedded Systems Engineering, Incheon National University, Incheon, South Korea

of variations in visual appearance, including shape, size, pose, cloth color, and body articulation, as shown in Fig. 1 The cluttered scenes and camera viewpoints, abrupt variations in motion, close interaction of objects, and occlusion are also important factors that might affect tracking algorithms' performance.

Several features and deep learning techniques have been proposed by researchers [16–19] which tried to overcome these challenges. The majority of developed techniques used frontal or side-view cameras, which might suffer from occlusion problems (phenomena occur when the person is obscure with another object or person, as shown in Fig. 1). To resolve this issue, some researchers [20] used multiple cameras in order to localize and detect people accurately, but at the cost of computation, installation expense, and transmission load, as each captured or recorded video sequences should be transmitted to the computer system for further processing.

In order to provide solution for occlusion problem shown in Fig. 1, many researchers e.g. [13, 21–24] suggested to utilize a single top view camera. As viewed from Fig. 2, occlusion problem in top view is much less likely to happen as compared to the front or side view, where inter object occlusion may occur when the scene becomes crowded. Because of this property top view based person tracking and detection got importance in many practical applications particularly in surveillance systems including person detection in outdoor and indoor environments, person counting in public areas [25–29], person tracking [29–34], action recognition, behavioural understanding [35], crowd analysis [36], person posture characterization [37] and industrial work flow [21, 33]. Using top view camera as well reduce the privacy issues because instead of face images the camera captures only person's body from top view [38]. Along with handling the occlusion problem, the single top view camera perspective may overcome computation, transmission load, power consumption, human resource, and installation expenses [39]. Figures 1 and 2 visualize the main difference between two different camera perspective. It can be observed from the sample images that perspective change in camera causes significant variation in the visual features (shape, size, pose, scale and body orientation) of a person.

In light of the above discussion, in this work, the framework consists of a top view 5G infrastructure, and deep learning-based detection and tracking module is presented. The framework consists of a top view IP camera that is utilized to record video streams. The video streams are transmitted to the server using the 5G infrastructure with high-speed and bandwidth. The server communicates these videos to the monitoring unit for further analysis and processing. The monitoring unit used deep learning models for multiple people tracking and detection. For multiple person detection, YOLOv3 (You Only Look Once) [41] has been applied, pre-trained on data set contains images recorded from frontal or normal view. In order to enhance the detection accuracy, transfer learning is performed by additional training of the existing detection model on top view person data set. As far as we know, this work might be considered the first effort

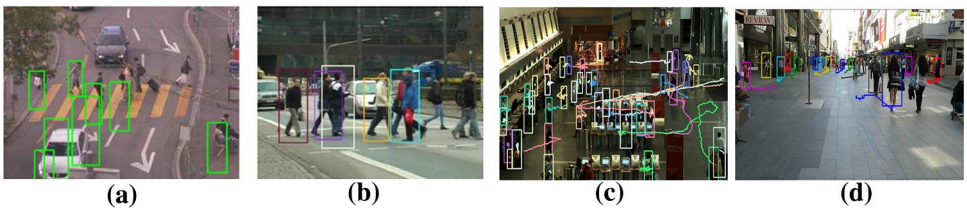

Fig. 1 Some example images taken from literature showing occlusion problem occurs in conventional frontal or side view. **a** Due to occlusion false detection occurs [12]. **b** Occlusion as person crossing each other [13]. **c** Occlusion caused by crowded scene [14]. **d** Occlusion because a side view camera perspective [15]

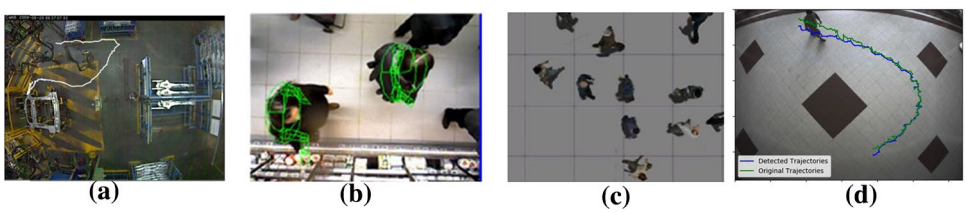

Fig. 2 Some sample images from literature showing how top view camera perspective overcome the problem of occlusion. **a** Top View Person tracking in industrial environment [22]. **b** Top view person tracking using (head and shoulder information) [13]. **c** Occlusion handling in top view [24]. **d** Top view person tracking in a wide field of view [40]

to apply transfer learning and training of object detection model for top view multiple person data set within 5G infrastructure. After having been detected, a person is tracked using an assigned unique tracking ID. For multiple people tracking, Deep SORT (Simple Online and Real-Time tracking) [42] algorithm is coupled with the detection model. The summary of the work is provided as:

- To provide a framework using 5G infrastructure for top view multiple people tracking. Deep learning-based tracking by detection method is employed, which includes detection by YOLOv3 and tracking by Deep SORT algorithm.
- To examine the performance of YOLOv3, pre-trained on data set containing frontal view sample images and tested on multiple person data set recorded using an IP camera, which is entirely distinct from the training data set.
- A transfer learning approach is adopted to enhance the detection accuracy of the model. The additional training is made utilizing top view data set embedded with the existing pre-trained model.
- To perform top view person tracking using a Deep SORT tracking algorithm coupled with a detection model.
- The detection model's performance is evaluated with and without transfer learning using top view multiple person data set (through accuracy) along with tracking results.

The work demonstrated in the paper is organized as follows: A summary of existing work in the area of top view person tracking and detection is elaborated in Sect. 2. The recorded multiple person data set using a single top view IP camera and utilized for testing and testing during experimentation is briefly discussed in Sect. 3. The framework used for detecting and tracking multiple persons within 5G infrastructure has been elaborated in Sect. 4. Section 5 illustrates a detailed explanation of experimental results. The performance evaluation of the model with and without transfer learning is also explained in Sect. 5. The conclusion of the presented work with possible future directions is presented in Sect. 6.

2 Literature review

Human tracking and detection in top view video sequences or images are considered one of the challenging tasks discussed in Sect. 1. Several top view person detection and tracking techniques are developed by researchers, based on a different traditional feature and few deep learning models. In top view images or video sequences, the person is detected and tracked using head information, head-shoulder information, or sometimes, the full top view person body information [43]. This section elaborates on a concise summary of various top view person tracking and detection techniques developed in recent years. The discussion is categorized in the traditional blob, feature, and deep learning-based techniques.

The majority person detection models developed for top view are based upon background subtraction and segmentation methods [25, 37, 44, 45]. Some researchers used features information such as [30] used shape information by considering the head of the person as cylindrical shape blob, similarly, [31] considered top view person shape as hemi ellipsoid. Ozturk et al. [46] assumed top view person body as an elliptical shape blob and performed detection in top view input images. Wu et al. [47] performed people tracking and [48] people detection using depth images. Furthermore, [49] adopted Hough circle, [38] used hair whorl shape information for person detection in top view images. Some researchers take advantage from color information [38, 46, 50, 51] while some used edge information's e.g canny edge detector or sobel filters [32, 46, 52] for top view person detection and tracking.

Many researchers used textual based information e.g. hair texture [38] for top view person detection. The region of interest (ROI) in a top view scene is also variable, some supposed only person head as (ROI) [32, 38] while other researchers considered head-shoulder information as ROI [45, 50, 53, 54] for person detection. Some also took the full person body as ROI for detection purposes [46, 53, 55, 56]. Along with person detection, top view person tracking methods are also developed by different researchers such as [27, 32, 50, 55, 57] used the Kalman filter for top view person tracking.

Likewise [31, 44, 46, 51, 57] utilized particle filter for person tracking in top view video sequences. In [25] authors utilized Hungarian algorithm for tracking people in top view frames. Moreover [51] considered median filter and [53] used mean shift algorithm for individual tracking. Burbano et al. [29] track people in top view and proposed graph structuring method. Bagaa et al. [58] provides an efficient tracking area framework using 5G networks. Several researchers take advantage of feature-based techniques such as in [59] developed an efficient top view human detection method using histogram of oriented gradient (HoG) features. In [34], authors used local ternary patterns features along with SVM classifiers for human detection and tracking.

Similarly, in [21] proposed a Rotated-HoG algorithm using top view industrial images for person detection. Another robust algorithm is proposed [22] using a wide-angle camera for recording top view video sequences of a person in indoor environments. The developed algorithm used variable size bounding boxes with different angles. Gao et al. [51] utilized HOG features for person detection and counting in complex environments. Lowe [60] and Ozturk et al. [46] used SIFT features to perceive the variation in a person's body and head. [34] used a fixed-sized detection

window along with the CoHOG feature-based technique for the detection of a person in images captured from the top view. Choi et al. [61] and Ertler et al. [26] developed a method based on various size detection windows that detect the person's body and head shoulder of the human in top view images. Choi et al. [61] used statistical measurement to reduce the extracted feature sized, like mean and standard deviation.

In [62], authors provide a top view feature-based method for person detection in an industrial and indoor environment. Likewise, in another work [33], proposed a feature-based approach for tracking multiple persons in an industrial environment using sample images recorded from the top view. Ullah et al. [40] employed blob based method and provided a rotation-invariant person tracking solution for top view surveillance. [63] also provides a comparison of some conventional tracking techniques using an overhead view person video sequences. The authors in [64] used an efficient rotated HoG based method along with SVM classifier for overhead view person detection. Few of them also used deep learning methods [26] for human tracking and detection using the fisheye camera. The majority of the proposed models' utilized the frontal view data set. Various researcher [24, 65–67] performed object detection and tracking tasks using aerial and satellite images . Correspondingly, some researchers applied deep learning techniques for top view object tracking and detection, but their work mainly for a single class object, mainly person [24, 68, 69]. Ahmed et al. [70] used two different deep learning models for top view multi-class object detection. Authors used Faster-RCNN and Mask-RCNN for top view object detection and segmentation, respectively. Ahmad et al. [71] used a convolutional neural network-based (CNN) tracking technique for multiple people tracking in an overhead view indoor and outdoor environments. In recent work [72] used top view person images and provided a comparison of three deep learning-based segmentation models.

3 Data set

The data set contains video sequences of multiple persons captured by utilizing a top view mounted IP camera. All sequences used in this work are captured in an indoor unconstrained university campus environment (Institute of Management Sciences, Pakistan.). The most prominent constituent elements are multiple people (commonly 2–20) dressed in different colored clothes, holding different objects in hands such as mobile phones, books, and bags. The video sequences are captured using Point Gray FlyCap2 camera with a Fujinon wide-angle lens installed approximately at the height of 4 m from the ground. The recording video frame rate is 20 fps, resolution 644×482, and a compression rate of 50%. As the wide-angle lens camera is used, which offers full coverage of the scene and solve the occlusion problem. The recorded duration of recorded video sequences is different. In each video sequence, the person is freely moving within the scene. For transfer learning, we used 1000 image patches extracted from different video sequences containing multiple people having variation in the size, appearance, pose, scale, and angle, as shown in sample frames discussed in Sect. 5. From the sample frames, it can be easily examined that person's visual appearance from a top view perspective is significantly different from the normal or frontal view. Non of these training images or video frames are used in any test set, for both training and testing different video sequences are used.

4 System overview

As discussed earlier, traditional surveillance systems either depend on complicated cables with dedicated software or fully-centralized processing, leading to costly solutions, longer analysis time, and huge bandwidth requirements. On the other hand, IP-based cameras enable with 5G connections reduce network load and support video streams at high bandwidth. In this work, the top view multiple people tracking framework using 5G infrastructure shown in Fig. 3 has been discussed. The overall framework is consists of a top view IP camera that is utilized to record video streams. The video streams are transmitted to the server using 5G infrastructure shown in Fig. 3. The 5G infrastructure transmit recorded video streams with high-speed and bandwidth, enhance deep learning technique, which requires intense processing to deliver a near-real-time solution for top view multiple people tracking. The server communicates these videos to the monitoring unit for further analysis and processing. The monitoring unit used a deep learning-based technique for multiple people tracking and detection. For multiple person detection, YOLOv3 [41] has been practiced, while for tracking, Deep SORT [42] algorithm is coupled with the detection module. The detail of deep learning-based

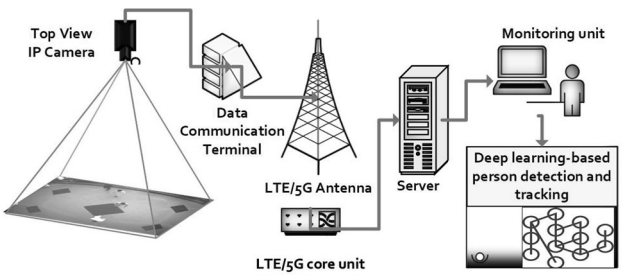

Fig. 3 Top view multiple people tracking using 5G infrastructure

top view multiple people tracking by detection is provided in the subsequent section.

4.1 Top view multiple people tracking using deep SORT and YOLOv3 with transfer learning

In this section, the top view multiple people tracking by detection framework shown in Fig. 4 has been discussed. The overall framework is split into two modules, the first one is the deep learning-based person detection module, and the second one is used for person tracking. In order to track the person in a top view video sequence, it must first be detected. There are number of object detection models available such as [73–78], but here the YOLOv3 [41] is employed which gives best results for generic object detection. This model utilized one stage network architecture in order to predict the class probabilities and corresponding detected bounding boxes. The original YOLOv3 model was trained on the COCO data set [79]. To improve the detection performance, transfer learning is adopted by additional training of model on top view, multiple person data set, and appended with a frontal view pre-trained model.

Furthermore, the person detection module is coupled with Deep SORT [42] tracking algorithm, which aims to identify and track multiple persons in the top view scene. Figure 4 shows the overall process practiced for the top view person tracking. The top view video streams are first converted into sample frames and fed into the YOLOv3 model, after detecting the person bounding box, the information is further processed by the Deep SORT tracking algorithm, which helps to track people in the scene. The deep learning-based tracking framework's detail applied for top view multiple people tracking by detection is elaborated in the following sections.

4.1.1 Person detection using YOLOv3 with transfer learning

In this work, YOLOv3 [41] is used for multiple person detection in top view scenes, as represented in Fig. 5. The module is categorized into two parts. The first part comprises of the pre-trained model with the COCO data set [79]. While in the second part, in order to further enhance the detection strength of the model for the top view person, transfer learning is adopted. It is widely practiced in various machine learning problems; it emphasized storing the knowledge developed during problem-solving and utilizing it for different but related problems solutions [80]. It is considered an effective key technology in deep learning, where models are trained on different data sets containing thousands of images. The main objective of the transfer learning approach is not to discard valuable information from the existing model and utilize existing trained model knowledge with a newly trained layer in order to solve new problems. Combining the existing and newly learned knowledge, makes or enables the models to provide a better and faster solution for different problems. The essential and key tool utilized in transferring learning is fine-tuning. In a variety of object detection and classification tasks, experiments utilized this tool as the lack of available data set of specific applications. In this work, the detection model YOLOv3 is additionally trained on top view multiple person data set and further embedded this newly trained model with frontal view pre-trained model as illustrated in Fig. 5, both weights files are combined, and a new learned model is generated which significantly improves the top view person detection results.

The YOLOv3 model used a single network architecture to predict the class probability and corresponding bounding boxes for the whole input image. The original YOLO [81] model contains twenty-four convolutional and two

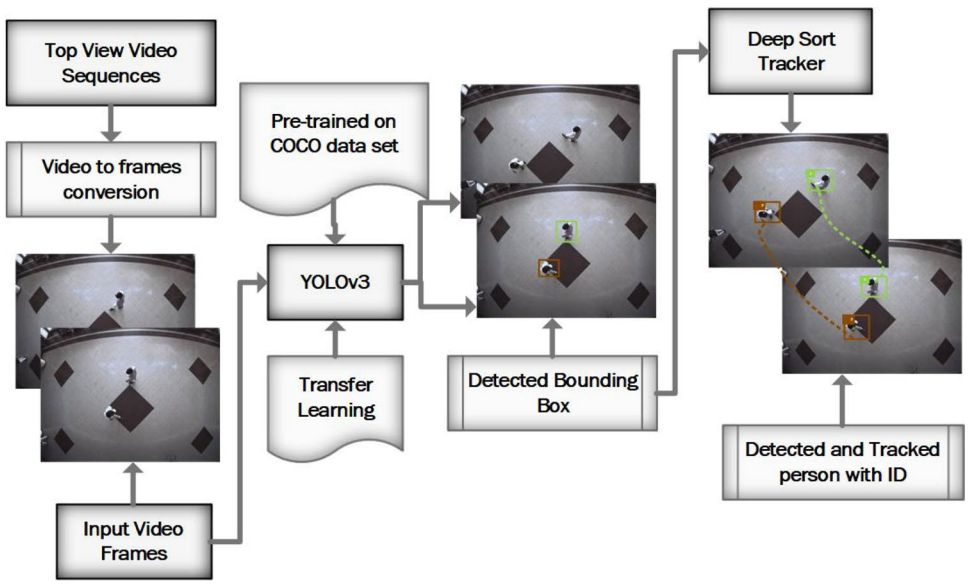

Fig. 4 The general deep learning based tracking by detection framework of top view person tracking. YOLOv3 model and Deep SORT algorithm is used for person detection and tracking respectively

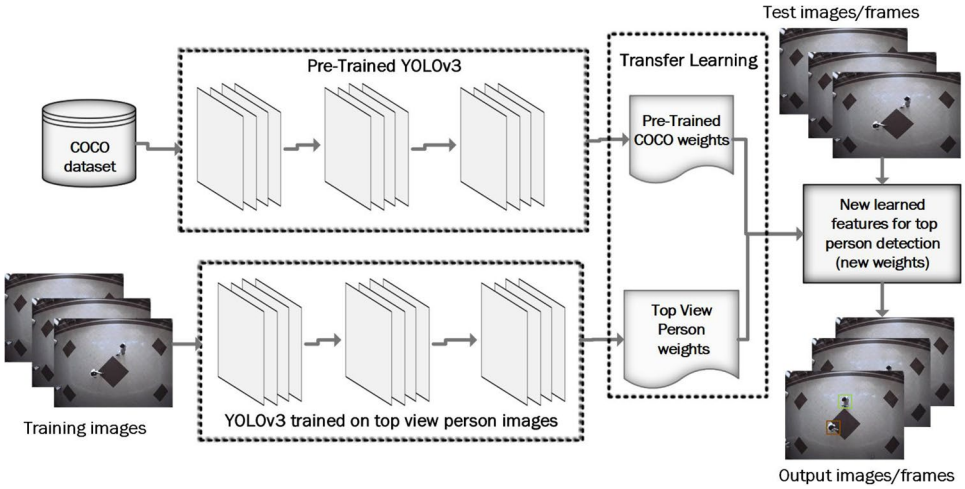

Fig. 5 The overall framework of top view multiple people detection using YOLOv3 with Transfer Learning. The detection model YOLOv3 pre-trained (utilizing frontal or normal view data set) is appended with the additional trained model on multiple person data set recorded using top view)

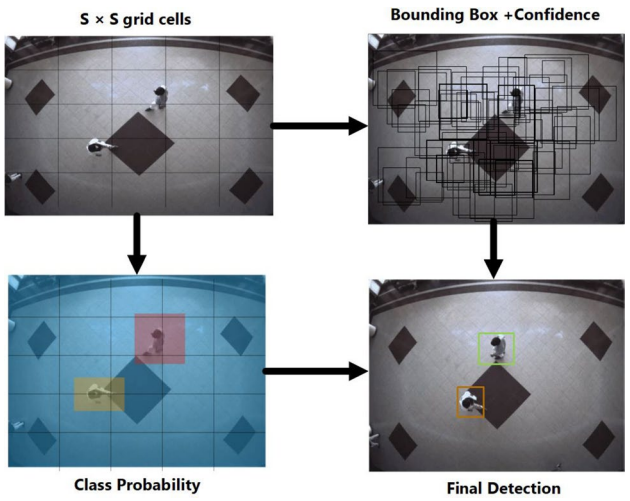

Fig. 6 Top view person detection approach using YOLO [81]

fully connected layers. The convolution layers are used for the extraction of the image while fully connected layers calculate the class prediction and probability. During person detection, the model, as visualized in Fig. 6, divides the input video frame into $S \times S$ regions, also called grid cells. These grid cells are associated with bounding box prediction and class probabilities. Each cell predicts the probability of whether the center of the person lies in the grid cell or not. If the prediction is positive, then bounding boxes and the confidence value for each positive detection is predicted. The confidence value signifies the degree of the detected bounding box as a person and defined as:

$$Conf(person) = Pr(person) \times IOU(Pred, Truth), \tag{1}$$

where $Pr(person)$ shows whether person is present in bounding box predicted or not (yes for 1, no for 0), and $IoU(Pred, Truth)$ is used for intersection of the predicted and real bounding box. It is given as [41]:

$$IoU(Pred, Truth) = area \frac{BoxT \cap BoxP}{BoxT \cup BoxP}. \tag{2}$$

In the above equation, $BoxT$ represents the ground truth box in the training set (manually labeled), and the predicted bounding box is expressed with $BoxP$. The area of intersection is represented by $area$.

For top view person detection, the suitable region is selected and predicted. After prediction, a confidence value is used to obtain the desired bounding box. Five values are predicted for all bounding boxes, including h, w, x, y, and confidence value, where width and height are represented by w, h, and bounding box center coordinates are represented by x, y. The low score confidence value is discarded by defining a threshold value, and the remaining multiple high confidence bounding boxes are processed, and final location parameters are derived by using non-maximal suppression. Finally, for the detected bounding box, the loss function is calculated. In the original work [81], loss function is the sum of regression loss and classification loss. However, in this work, only one object, i.e., person, is considered. Therefore, the loss function for this work is given as:

$$loss(person) = L_c + L_{IoU} \tag{3}$$

where L_c represents predicted bounding box coordinates loss and L_{IoU} is used for calculating the ground truth bounding box coordinates loss.

The size of the person in top view image is different, YOLO v3 loss function calculates same loss for all bounding boxes. However, the effect of small and large objects on loss function are different for the whole image. Thus, in order to improve the loss function of the coordinates contrast

normalization is used [82] . The loss function of coordinate L_c is given by [81]:

$$\mathscr{L}_c = \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{person} \Big[(x_i - x_i^*)^2 + (y_i - y_i^*)^2 + (\sqrt{w_i} - \sqrt{w_i^*})^2 + (\sqrt{h_i} - \sqrt{h_i^*})^2 \Big]. \tag{4}$$

In the above equations λ_{coord} is the scale parameters used for bounding box coordinates predictions here $\lambda_{coord} = 5$ ([81]) . x_i, y_i, h_i, w_i are the predicted positions of detected bounding box in i_{th} cell, while $x_i^*, y_i^*, h_i^*, w_i^*$ are the actual positions of bounding box in the i_{th} cell. The above equation calculates the loss function related to the predicted bounding box having coordinates value x, y. The I_{ij}^{person} which shows the possibility of the detected person in the j_{th} bounding box. The term λ is constant, the above function calculates sum over each bounding box, using ($j = 0$ to B) as predictor for each grid cell ($i = 0$ to S^2). The function is defined as [81],

The L_{IOU} of IOU_{pred}^{truth} is calculated as:

$$\mathscr{L}_{IoU} = \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{person} [(\xi_i - \xi_i^*)]^2 + \lambda_{no-person} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{no-person} [(\xi_i - \xi_i^*)]^2. \tag{5}$$

In Eq. 5 $\lambda_{no-person}$ shows classification error; similarly ξ_i and ξ_i^* is the confidence value of the i_{th} grid cell in the predicted and original sliding window. I_{person} represents whether the person is detected in j_{th} bounding box of grid cell i or not. In case the target person is present in j_{th} bounding box and i_{th} grid cell, then the function is equal to 1; otherwise, it becomes 0.

4.1.2 Top view person tracking using deep SORT

For tracking a person from the top view, deep learning-based tracking algorithm Deep SORT [42] is used as shown in Fig. 7. It is mainly based on the frame-by-frame data association method and Kalman filtering. The filtering is used to assess the existing tracks in current video frames. It usually used $x', y', h', \gamma', u, v, h, \gamma$, where x', y', h', γ' is tracking velocity of each coordinate of detected bounding box and (u, v, h, γ) is the positions of bounding box [42].

The Kalman filter is employed using Deep SORT using linear observation and constant velocity. Consequently, each existing track's position in the prevailing frame has been estimated in the next successive frame. The track estimation is made by using spatial information of the bounding box. To obtain the appearance information of each detection and tracking, feature extraction is performed using an appearance descriptor. This descriptor is trained using the CNN

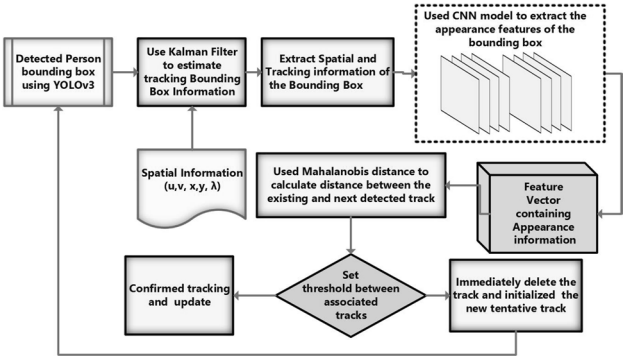

Fig. 7 General frame work of top view person tracking using deep SORT [42]

model. Here, the features are extracted using the trained model, and a feature vector is formed. The vector places the same identity features together, and features of unique identities are placed distant from each other.

Using the information extracted from the appearance descriptor, the new detection results can now be associated with present tracking results in the next successive frame. For that purpose, a detection threshold is defined so that the low detections' results are not considered. In the next successive frame, each detection result is now associated using a threshold. The cost matrix is used by Deep SORT algorithm to represent the appearance, and spatial similarities between the new detections and tracks using two distance values as represented follow [42]:

$$d^{(1)}(i,j) = (d_j - y_i)^T S_i^- 1 (d_j - y_i). \tag{6}$$

In the above equation, y_i, S_i represents the i_{th} projection track is measurement space and for j_{th} new detection d_j is used. It is also called the Mahalanobis distance, which is calculated difference between new detection j_{th} and estimated position i_{th} track. Further, using the above metric, unlikely associations is excluded by using the Mahalanobis distance threshold between i_{th} track and j_{th} detection given as [42]:

$$b_{i,j}^{(1)} = 1[d^{(1)}(i,j) < t^{(1)}]. \tag{7}$$

For estimating second distance value which represents the appearance information the following equation [42] is used. This second distance value estimated the smallest cosine distance between the j_{th} detection and i_{th} track as follows:

$$d^{(2)}(i,j) = min\Big(1 - r_j^T r_k^{(i)} | r_k^{(i)} ER_i\Big). \tag{8}$$

In the above r represents appearance descriptor and R_i is used to represents the appearance of at least 100 objects (persons) in the i_{th} track. To set the threshold between the association tracks we used [42]:

$$b_{i,j}^{(1)} = 1[d^{(2)}(i,j) < t^{(2)}]. \tag{9}$$

If distance value is small it is equal to 1 if distance value is small and 0 if distance is large. For more details we refer readers [42]. We combine the above cost functions using the below matrix:

$$c_{i,j} = \lambda d_{(1)}(i,j) + (1-\lambda) d_{(2)}(I,j). \tag{10}$$

For matching the spatial and appearance information the gate function is given as [42]:

$$b_{i,j} = \prod_{m=1}^{2} b_{i,j}^{m}. \tag{11}$$

If the above equation's value is equal to 1, it indicates that both appearance and spatial gate functions are equal to 1 if 0 if not equal. It also indicates (i, j) is a true match between appearance and spatial information. Therefore, in every new video frame, the detections and tracks are associated with utilizing the above cost and gate functions. For processing, tracking in the video sequence, in the next new video frame in the case when the new detection is effectively associated with the present track, then tracking has continued. While if not associated or matched, it is set to zero. Therefore, in that case, the new detections fail. In such a case, new detections fail to associate in frame f with the existing; then, new detection is initialized as tentative tracks. Therefore, Deep SORT algorithm continuously verifies and associated it with new detections in next $(f+1), (f+2), \ldots (f+t)$ tentative frames. As long as it is successfully associated, then that track is confirmed for tracking and updated. Otherwise, it is deleted immediately.

5 Experiments, results and discussion

This section provides detail of different experiments performed in this work. Several top view person video sequences are used containing a number of people having variation is appearances, heights, poses, scales, shapes, illumination conditions, angle, cameras resolution, and aspect ratio. Both of the models have been implemented using OpenCV. Python programming language with PyTorch framework is practiced for performing experimentation steps. The detection results are divided into two phases; first, the YOLOv3 pre-trained (COCO data set) weights file is used for testing, while in the second phase YOLOv3 detection model is trained on top view multiple person data set. Once training is completed, both weights files are combined using transfer learning, and testing is performed on different video sequences. This section is categorized into three sections. The first section discusses the detection and tracking results using the pre-trained YOLOv3 model along with the Deep SORT tracking algorithm. As discussed earlier, to enhance/improve the detection model's performance, transfer learning is adopted, which has been discussed in the second section. The third section discusses different parameters used for performance evaluation of the detection model and tracking algorithm.

5.1 Results top view person detection and tracking using pre-trained YOLOv3 and deep SORT

The testing results of the pre-trained YOLOv3 model using the COCO data set [79] can be visualized from Figs. 8, 9 and 10. The pre-trained model is tested on different video sequences. In the first video sequence, two people are freely moving in the top view scene; it can be noted from

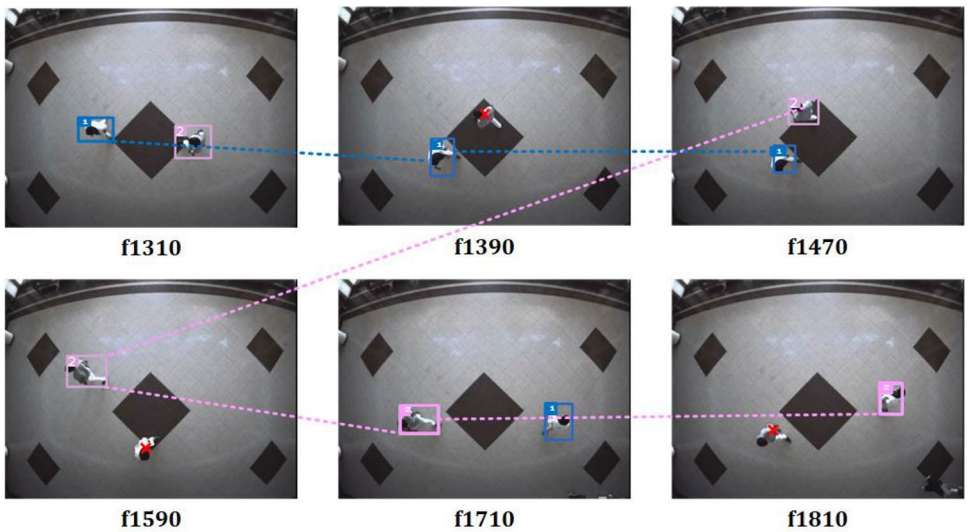

Fig. 8 Detection and tracking results of pre-trained YOLOv3 and Deep SORT tracking algorithm for top view person video sequence. The visualization results can be seen for the initial few frames; two persons are allowed to walk within the top view scene freely in this video sequence

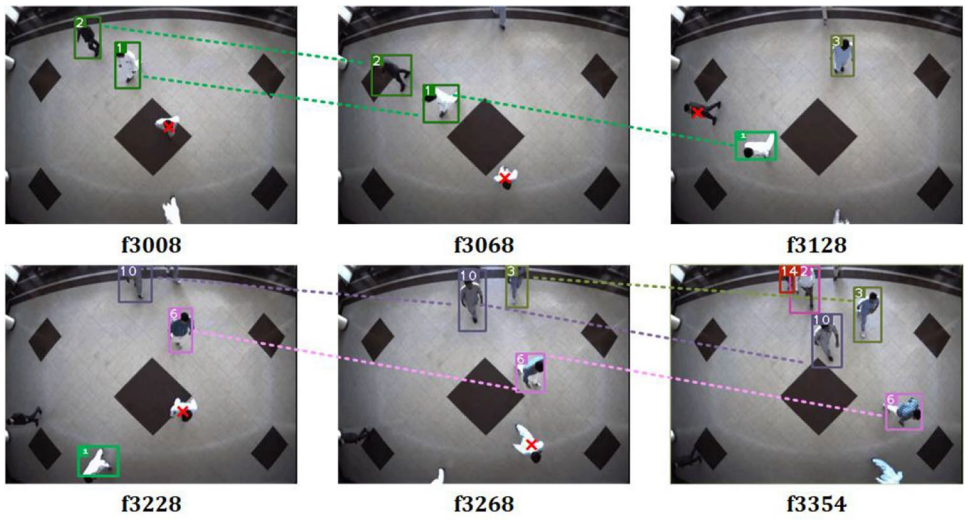

Fig. 9 Detection and tracking results of pre-trained YOLOv3 and Deep SORT tracking algorithm for top view multiple person video sequence (the results are shown for few sample frames)

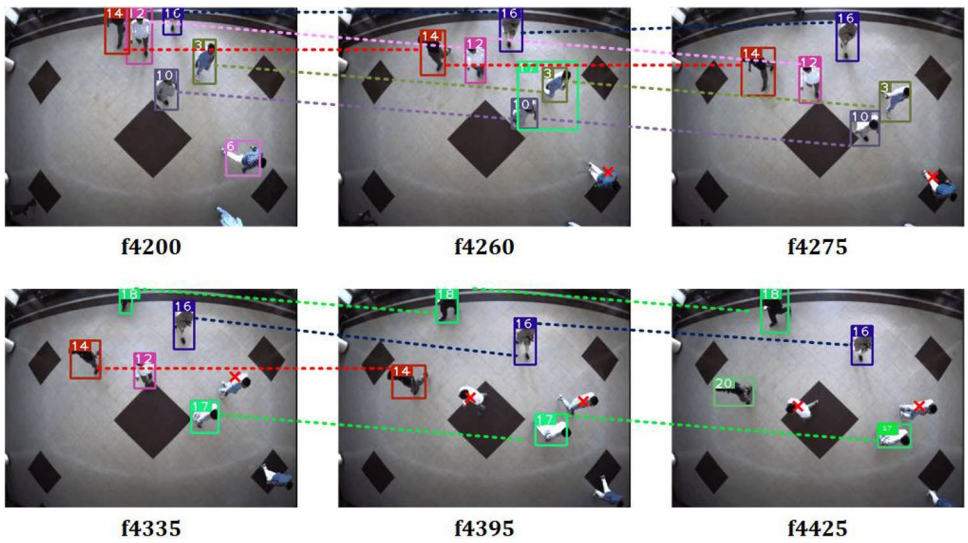

Fig. 10 Detection and tracking results of pre-trained YOLOv3 and Deep SORT tracking algorithm for top view multiple person video sequence. In this video sequence the number of multiple people are entering in the scene is increasing (the results are shown for few sample frames)

the sample frames, that the person's visual appearance is not similar as in frontal or side view. At the different locations in the scene, the size of the person is varying Fig. 8. Thus, the pre-trained detection model is not capable of detecting the person, leading to detection and tracking failure. Although the model still shows better detection and tracking results. The good results are demonstrated with detected person bounding boxes while not detected results (false positives) are manually marked with a red cross. It can be seen from the sample frames that person1 and person2 are detected and tracked in some frames, but in some cases where there is a change in the appearance of the person appearance, the models give wrong results (false positives). This happens maybe because the pre-model trained model is used so that the person's appearance might be confusing for the detection model.

Similarly, we also tested the detection and tracking algorithms for another top view video sequence where multiple people are entering the top view scene. The overall testing performance of the tracking algorithm and detection model is good, but in some cases, the person in the top view is not detected and tracked. Also, in some cases, the tracking ID is swapped to another person, which shows that model assigned the same tracking ID to a different person. In Fig. 9, the visualization results of the pre-trained model for multiple persons entering in top view scene video sequence can be seen. From the sample frame (f3008, f3068), two-person entering the scene are correctly detected and tracked, but one person marked with the red cross is not detected and tracked. As the person moves at the center and the lower body is obscure, maybe that is why the pre-trained model gives false positives and cannot detect it. Similarly, in the sample frame

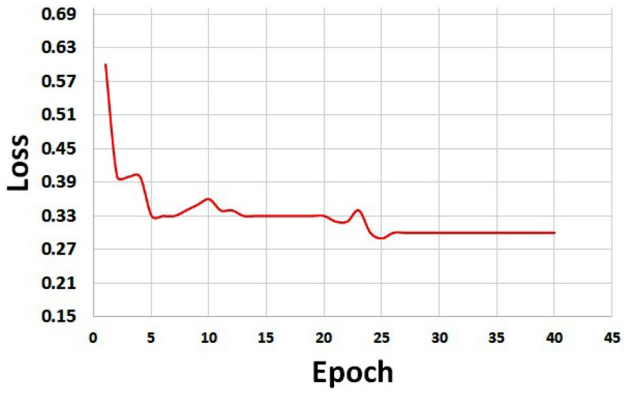

Fig. 11 Training loss of YOLOv3 using top view person data set

(f3008, f3068, f3228), the person1 is not detected because of sudden illumination changes. In the sample frame, Fig. 9 (f3228, f3268,f3354), multiple people entering the scene are accurately detected and tracked, revealing the effectiveness of deep learning. In sequence, the number of people is varying as seen from the sample frame (f3354); the deep learning models detect and tracked up to fourteen people in the top view scene.

The testing results for another video sequence in which multiple people are entering the scene can also be visualized in Fig. 10. From sample frames, it can be noticed that deep learning models' detection and tracking results are promising. Although the detection model is pre-trained using data set containing frontal view images, it still gives good results for the top view data set. In some regions in the top view scene, the model cannot track and detect the person, which are manually marked using the red cross in sample frames. For example, in Fig. 10 (f4260) the detection model assigned the same bounding box to too people the reason may be the person are walking two close to each other or maybe the shadow of the persons on the floor. Similarly, in sample frames (f4335, f4395, f4425), the detection and tracking model lost the ID of person14, which may be due to change in the person's appearance.

5.2 Results of top view person detection and tracking using YOLOv3 and deep SORT after transfer learning

As discussed earlier, to enhance/improve the person detection accuracy of YOLOv3 for top view data set, the transfer learning approach is applied. The YOLOv3 model is additionally trained using top view multiple person data set. Over 10,000 sample frames/images of multiple persons were collected in an unconstrained indoor environment, as discussed in Sect. 3. For testing and training, the data set video frames are split into 30% and 70%, respectively. For training, batch size of 64 and the number of epochs = 40 is utilized. The training loss and training accuracy are obtained at the end of the 40th epoch, as depicted in Figs. 11 and 12, respectively. After training the model, a new weight file is generated; with the help of transfer learning, weights file is combined with a pre-trained weights file (COCO data set).

The top view video sequences are tested using the newly learned model. The experimental results illustrate that transfer learning significantly improves results for top view data set. Figures 13, 14 and 15 demonstrates the output results of the top view multiple people tracking and detection framework after applying transfer learning.

We tested the model on the same video sequences as discussed in the above section. The first video sequence used for testing is mainly covered indoor top view environment containing two peoples. The movement of the person within the scene is not restricted. Both people are moving freely in the scene; the person's appearance is affected due to movement with respect to camera position or radial distance. It can be observed from Fig. 13 that the visual features of the person in all sample frames are not similar. However, as the detection model is additionally trained on top view, multiple person data set and combined with the pre-trained weights file, so it efficiently detects and tracks both of the people. The model also assigns the tracking ID to each person, as depicted in Fig. 13. Even there is a significant variation in the person's appearance, for example, in frames (f0010, f0574), the movement of people across the scene and their location with respect to camera position is different and scale is varying. Even then, the detection model with transfer learning detect and track multiple persons without any failure and also adjust the bounding boxes according to their sizes.

The YOLOv3 top view multiple person data set trained model is also tested for another video sequence in which multiple persons are entering in the top view scene, as shown first sample frames of Fig. 14. It can be seen from

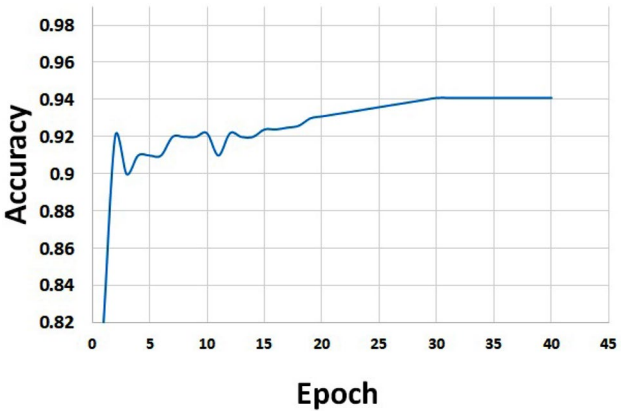

Fig. 12 Training accuracy of YOLOv3 using top view person data set

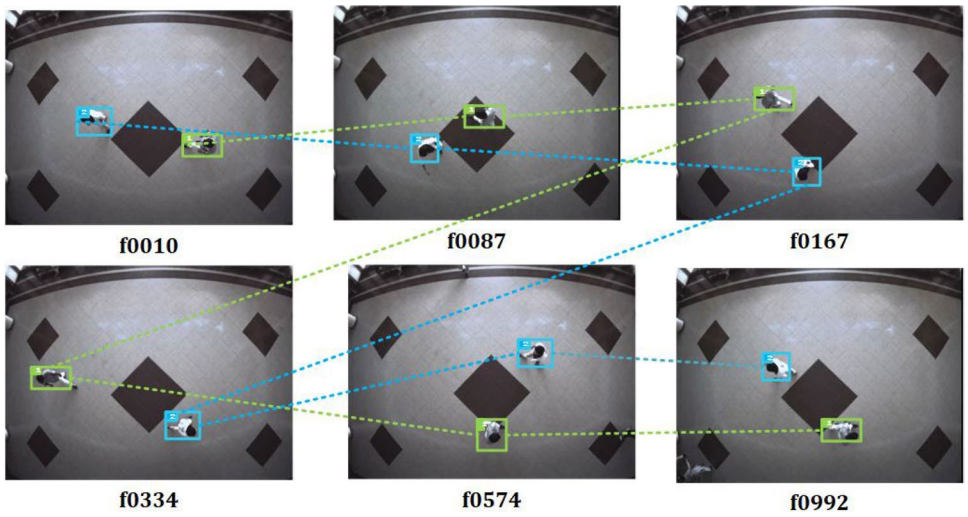

Fig. 13 Testing results for top view person video sequence of Deep SORT and YOLOv3 after transfer learning. The visualization results can be seen for the initial few frames, in this video sequence two persons are allowed to walk freely within the top view scene

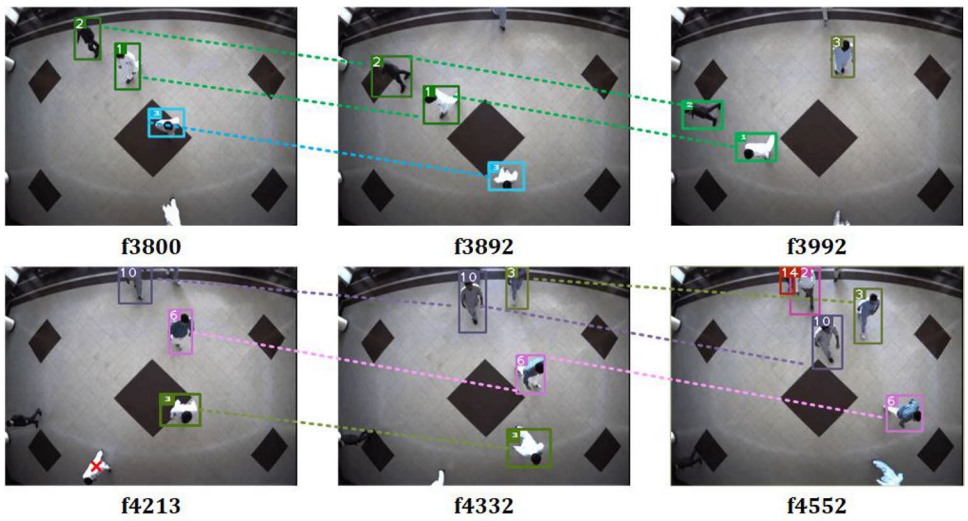

Fig. 14 Testing results for top view multiple person video sequence of Deep SORT and YOLOv3 after transfer learning. In this video sequence multiple people are entering in the scene

sample frames that after applying transfer learning, the accuracy of the detection model is significantly enhanced, the model efficiently detects, counts, and tracks multiple people in the top view scene. In Figs. 14 and 15, due to space limitations, the results are shown for a few sample frames.

In Fig. 15, the trained model's testing results for multiple person sample frames are depicted. In sample frames person are closely walking in the scene, wearing different color cloths. It can be visualized from the sample frames, although people are closely interacting with each other; the deep learning model still discriminates each person in the top view scene. In some places, although the model has not detected the person as compared to pre-trained mode, the accuracy is significantly improved.

5.3 Performance evaluation

In this work, different quantitative parameters are used for evaluating the performance of deep learning-based top view multiple person tracking framework. For evaluating the detection model's performance, YOLOv3 precession, recall, and accuracy have been used. While for tracking, we used the same parameter as [83]. The precision, recall and accuracy is given as:

$$Precision = \frac{tp}{tp + fp} \tag{12}$$

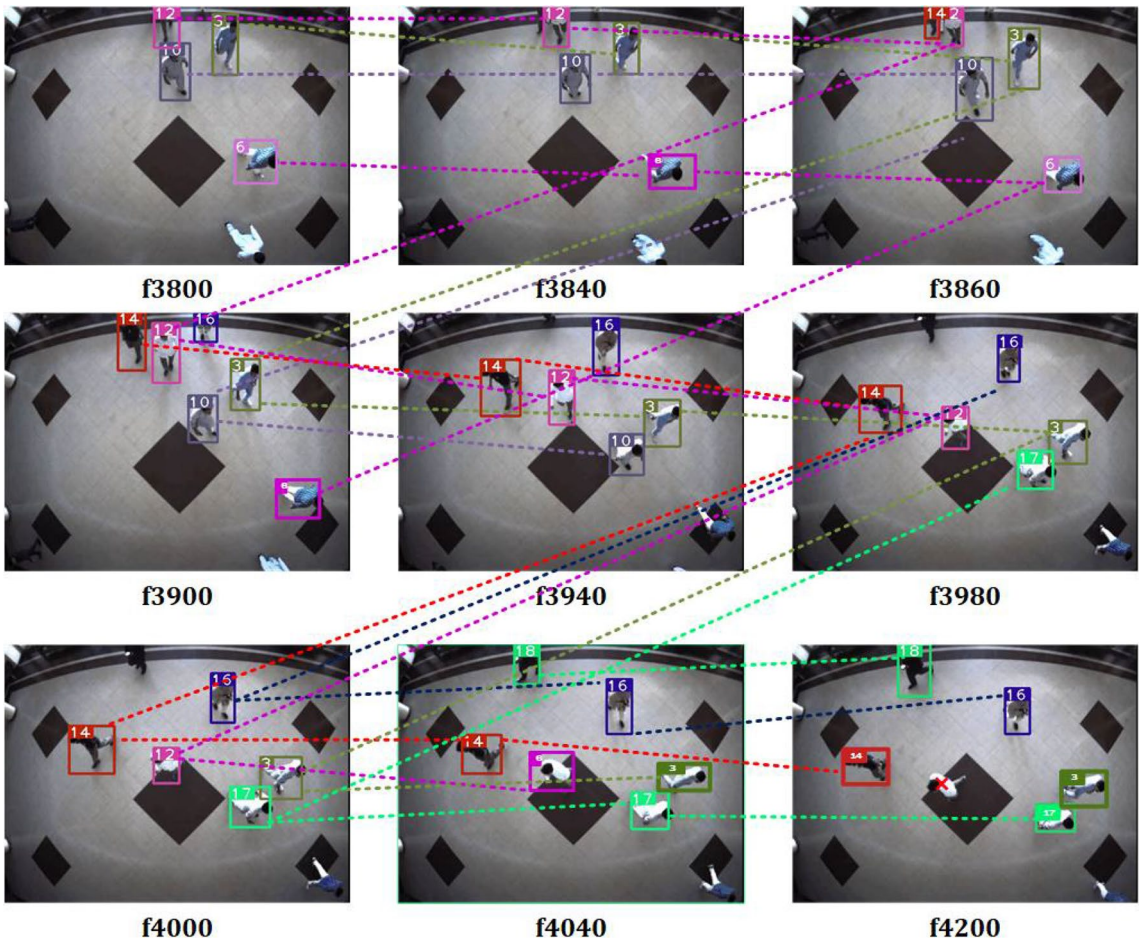

Fig. 15 Testing results for top view multiple person video sequence using Deep SORT and YOLOv3 after transfer learning. In this video sequence the number of multiple persons are entering in the scene is increasing

$$Recall = \frac{tp}{tp + fn} \tag{13}$$

$$Accuracy = \frac{tp + tn}{tp + fp + tn + fn}. \tag{14}$$

In above Eqs. (12)–(14), *tp* represents the number of true positive, bounding boxes correctly detected as the person in top view video sequences. *fp* shows the number of false-positive, bounding boxes incorrectly detected as a person, *tn* true negative shows the number of bounding boxes correctly recognized as background and *fn* false-negative shows when model incorrectly recognized background as a person. Results of precision, recall, and accuracy for top view person detection is shown in Fig. 16. It can be analyzed, that when the model is additionally trained for top view data set.

To estimate the performance of the tracking algorithm, the most widely metric MOTA and MOTP are used in this work as in original work [42] and [83, 84]. The main

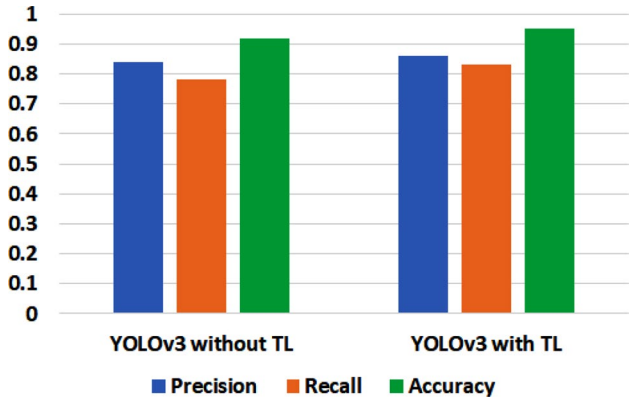

Fig. 16 Precision, recall and accuracy of pre-trained YOLOv3 without transfer learning and YOLOv3 trained with transfer learning for top view multiple person detection

Table 1 Evaluation results of tracking algorithm

S.NO	Deep SORT with YOLOv3 (without transfer learning) (%)	Deep SORT with YOLOv3 (with transfer learning) (%)
MOTA	92	96
MOTP	90	95

reason for using these parameters is because it combines three different sources of errors, such as [84].

$$MOTA = 1 - \frac{\sum_t (fn_t + fp_t + idsw_t)}{\sum_t g_t}. \quad (15)$$

It shows multiple object tracking accuracy (MOTA) in terms of the miss rate or the number of mistakes the tracking algorithm makes in terms of false positives, miss-matches, and failures to recover tracks. The t in the above equation represents frame index gt is the number of ground truth objects or person in our case. In order to represent how well the exact position of the person in the top view scene is estimated, multiple objects tracking precision (MOTP) is used given as [84]:

$$MOTP = \frac{\sum t, id_{t,i}}{\sum_t c_t} \quad (16)$$

In the above equation, the number of matches in the video frame t is denoted by c_t and bounding box overlap of the target object, i.e., person i with the ground truth box is represented as $d_{t,i}$. The tracking accuracy for the top view person data set is shown in Table 1. It shows MOTA and MOTP of Deep SORT tracking algorithm using the YOLOv3 detection model with and without transfer learning.

6 Conclusion

In this work, for top view, multiple people tracking, deep learning-based tracking by detection framework is proposed using 5G infrastructure. For detection purposes, the YOLOv3 detection model is used. Since the detection model was pre-trained using data set containing frontal or side view images and substantial variations in appearances, visibility, size, shape, and pose of the person in the top view scene, it still gives encouraging results. To further enhance the detection accuracy of the pre-trained model YOLOv3, transfer learning is adopted. The detection model is additionally trained using top view multiple person data set. The newly trained model weight file is combined with a pre-trained model, i.e., COCO weights file. For a top view, multiple people tracking the detection model is coupled with the Deep SORT tracking algorithm. As far as we know, this work is the first effort that used transfer learning and trained the YOLOv3 (object detection model) on top view multiple person data set the training dataset containing multiple persons, having variation in poses, scales, sizes, and appearances. The experimental results illustrate the robustness and efficiency of the top view deep learning-based person detection model. With pre-trained weights file without transfer learning, the accuracy of the detection model is 92%, and with transfer learning, the model achieves an accuracy of 95%. For top view person tracking, the model achieves tracking accuracy of 96%.

In future work, the model may be extended for multiple top view object data set, by using transfer learning, the model might be additionally trained for different top view objects with different backgrounds and scenes. The detection accuracy of model might be further improved by training the model on a completely top view multiple object data set.

Acknowledgement This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (No. 2018045330).

References

1. Jang YM, Cano JC, Yang K, and Choi Y-J (2016) Enabling technologies towards next generation. Mobile Syst Netw 2016:9805636. https://doi.org/10.1155/2016/9805636
2. Chen Y, Yang X, Zhong B, Pan S, Chen D, Zhang H (2016) Cnn-tracker: online discriminative object tracking via deep convolutional neural network. Appl Soft Comput 38:1088–1098
3. Zhan B, Monekosso DN, Remagnino P, Velastin SA, Xu L-Q (2008) Crowd analysis: a survey. Mach Vis Appl 19(5–6):345–357
4. Wu X, Huang G, Sun L et al (2016) Fast visual identification and location algorithm for industrial sorting robots based on deep learning. Robot 38(6):711–719
5. Clift LG, Lepley J, Hagras H, Clark AF (2018) Autonomous computational intelligence-based behaviour recognition in security and surveillance. In: Counterterrorism, crime fighting, forensics, and surveillance technologies II, vol 10802. International Society for Optics and Photonics. SPIE, pp 173–179. https://doi.org/10.1117/12.2325577
6. Hodgetts HM, Vachon F, Chamberland C, Tremblay S (2017) See no evil: cognitive challenges of security surveillance and monitoring. J Appl Res Mem Cogn 6(3):230–243
7. Jeong Y, Son S, Jeong E, Lee B (2018) An integrated self-diagnosis system for an autonomous vehicle based on an IOT gateway and deep learning. Appl Sci 8(7):1164
8. Bansal P, Kockelman KM (2018) Are we ready to embrace connected and self-driving vehicles? A case study of texans. Transportation 45(2):641–675
9. Verschae R, Ruiz-del Solar J (2015) Object detection: current and future directions. Front Robot AI 2:29
10. Borji A, Cheng M-M, Jiang H, Li J (2015) Salient object detection: a benchmark. IEEE Trans Image Process 24(12):5706–5722
11. Haghighat M, Abdel-Mottaleb M (2017) Low resolution face recognition in surveillance systems using discriminant correlation analysis. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE, pp 912–917

12. Breitenstein MD, Reichlin F, Leibe B, Koller-Meier E, Van Gool L (2011) Online multiperson tracking-by-detection from a single, uncalibrated camera. IEEE Trans Pattern Anal Mach Intell 33(9):1820–1833
13. Choi J-W, Moon D, Yoo J-H (2015) Robust multi-person tracking for real-time intelligent video surveillance. ETRI J 37(3):551–561
14. Shu G, Dehghan A, Oreifej O, Hand E, Shah M (2012) Part-based multiple-person tracking with partial occlusion handling. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp 1815–1821
15. Liu P, Li X, Liu H, Fu Z (2019) Online learned Siamese network with auto-encoding constraints for robust multi-object tracking. Electronics 8(6):595
16. Zou Z, Shi Z, Guo Y, Ye J (2019) Object detection in 20 years: a survey. arXiv:1905.05055
17. Yao R, Lin G, Xia S, Zhao J, Zhou Y (2019) Video object segmentation and tracking: a survey. arXiv:1904.09172
18. Zhou S, Ke M, Qiu J, Wang J (2018) A survey of multi-object video tracking algorithms. In: International Conference on Applications and Techniques in Cyber Security and Intelligence. Springer, New York, pp 351–369 **(ISBN: 978-3-319-98776-7)**
19. Li P, Wang D, Wang L, Lu H (2018) Deep visual tracking: review and experimental comparison. Pattern Recogn 76:323–338
20. Anuj L, Krishna MG (2017) Multiple camera based multiple object tracking under occlusion: a survey. In: 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA). IEEE, pp 432–437
21. Ahmed I, Carter JN (2012) A robust person detector for overhead views. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012). IEEE, pp 1483–1486
22. Ahmed I, Adnan A (2017) A robust algorithm for detecting people in overhead views. Clust Comput 21(1):1–22. https://doi.org/10.1007/s10586-017-0968-3
23. Ahmad M, Ahmed I, Ullah K, Khan I, Adnan A (2018) Robust background subtraction based person's counting from overhead view. In: 2018 9th IEEE annual ubiquitous computing, electronics mobile communication conference (UEMCON), pp 746–752
24. Migniot C, Ababsa F (2016) Hybrid 3D–2D human tracking in a top view. J Real Time Image Proc 11(4):769–784
25. Vera P, Monjaraz S, Salas J (2016) Counting pedestrians with a zenithal arrangement of depth cameras. Mach Vis Appl 27(2):303–315
26. Ertler C, Posseger H, Optiz M, Bischof H (2017) Pedestrian detection in RGB-D images from an elevated viewpoint. In: 22nd Computer Vision Winter Workshop. TU Wien, Pattern Recongition and Image Processing Group, Vienna
27. Kristoffersen M, Dueholm J, Gade R, Moeslund T (2016) Pedestrian counting with occlusion handling using stereo thermal cameras. Sensors 16(1):62
28. Malawski F (2014) Top-view people counting in public transportation using kinect. Chall Mod Technol 5(4):17–20
29. Burbano A, Bouaziz S, Vasiliu M (2015) 3D-sensing distributed embedded system for people tracking and counting. In: 2015 International Conference on Computational Science and Computational Intelligence (CSCI). IEEE, pp 470–475
30. Zhang Z, Venetianer PL, Lipton AJ (2008) A robust human detection and tracking system using a human-model-based camera calibration. In: The Eighth International Workshop on Visual Surveillance-VS2008. Marseille. https://hal.inria.fr/inria-00325644/file/VS2008-Poster-r.pdf
31. Tseng T-E, Liu A-S, Hsiao P-H, Huang C-M, Fu L-C (2014) 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. IEEE, pp 4077–4082
32. García J, Gardel A, Bravo I, Lázaro JL, Martínez M, Rodríguez D (2013) Directional people counter based on head tracking. IEEE Trans Ind Electron 60(9):3991–4000
33. Ahmed I, Ahmad A, Piccialli F, Sangaiah AK, Jeon G (2018) A robust features-based person tracker for overhead views in industrial environment. IEEE Internet of Things J 5(3):1598–1605
34. Rauter M (2013) Reliable human detection and tracking in top-view depth images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 529–534
35. Lin Q, Zhou C, Wang S, Xu X (2012) Human behavior understanding via top-view vision. AASRI Procedia 3:184–190
36. Ryan D, Denman S, Sridharan S, Fookes C (2015) An evaluation of crowd counting methods, features and regression models. Comput Vis Image Underst 130:1–17
37. Hsu T-W, Yang Y-H, Yeh T-H, Liu A-S, Fu L-C, Zeng Y-C (2016) Privacy free indoor action detection system using top-view depth camera based on key-poses. In: 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, pp 004058–004063
38. Nakatani R, Kouno D, Shimada K, Endo T (2012) A person identification method using a top-view head image from an overhead camera. JACIII 16(6):696–703
39. Ahmad M, Ahmed I, Ullah K, Khan I, Khattak A, Adnan A (2019) Energy efficient camera solution for video surveillance. Int J Adv Comput Sci Appl. https://doi.org/10.14569/IJACSA.2019.0100367
40. Ullah K, Ahmed I, Ahmad M, Rahman AU, Nawaz M, Adnan A (2019) Rotation invariant person tracker using top view. J Ambient Intell Human Comput:1–17 (**Springer**)
41. Redmon J, Farhadi A (2018) Yolov3: An incremental improvement. arXiv:1804.02767
42. Wojke N, Bewley A, Paulus D (2017) Simple online and realtime tracking with a deep association metric. In: IEEE international conference on image processing (ICIP). IEEE, pp 3645–3649
43. Ahmad M, Ahmed I, Ullah K, Ikhan , Khattak A, Adnan A (2019) Person detection from overhead view: a survey. Int J Adv Comput Sci Appl. https://doi.org/10.14569/IJACSA.2019.0100470
44. Iguernaissi R, Merad D, Drap P (2018) People counting based on kinect depth data. In: Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods—vol 1: ICPRAM. SciTePress, Setúbal, pp 364–370. https://doi.org/10.5220/0006585703640370
45. Perng J-W, Wang T-Y, Hsu Y-W, Wu B-F (2016) The design and implementation of a vision-based people counting system in buses. In: 2016 International Conference on System Science and Engineering (ICSSE). IEEE, pp 1–3
46. Ozturk O, Yamasaki T, Aizawa K (2009) Tracking of humans and estimation of body/head orientation from top-view single camera for visual focus of attention analysis. In: 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops. IEEE, pp 1020–1027
47. Wu C-J, Houben S, Marquardt N (2017) Eaglesense: tracking people and devices in interactive spaces using real-time top-view depth-sensing. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. ACM, New York, pp 3929–3942. https://doi.org/10.1145/3025453.3025562
48. Wetzel J, Laubenheimer A, Heizmann M (2020) Joint probabilistic people detection in overlapping depth images. IEEE Access 8:28Ã‚Â 349–28Ã‚Â 359
49. Van Oosterhout T, Bakkes S, Kröse BJ et al (2011) Head detection in stereo data for people counting and segmentation. In: VISAPP, pp 620–625
50. Wateosot C, Suvonvorn N et al (2013) Top-view based people counting using mixture of depth and color information. In:

The second Asian conference on information systems. ACIS **(Citeseer)**
51. Gao C, Liu J, Feng Q, Lv J (2016) People-flow counting in complex environments by combining depth and color information. Multimedia Tools Appl 75(15):9315–9331
52. Mukherjee S, Saha B, Jamal I, Leclerc R, Ray N (2011) Anovel framework for automatic passenger counting. In: 2011 18th IEEE International Conference on Image Processing. IEEE, pp 2969–2972
53. Velipasalar S, Tian Y-L, Hampapur A (2006) Automatic counting of interacting people by using a single uncalibrated camera. In: 2006 IEEE International Conference on Multimedia and Expo. IEEE, pp 1265–1268
54. Yu S, Chen X, Sun W, Xie D (2008) A robust method for detecting and counting people. In: 2008 International Conference on Audio, Language and Image Processing. IEEE, pp 1545–1549
55. Yahiaoui T, Meurie C, Khoudour L, Cabestaing F (2008) A people counting system based on dense and close stereovision. In: International Conference on Image and Signal Processing. Springer, Berlin, Heidelberg, pp 59–66 (**ISBN: 978-3-540-69905-7**)
56. Cao J, Sun L, Odoom MG, Luan F, Song X (2016) Counting people by using a single camera without calibration. In: Chinese control and decision conference (CCDC). IEEE, pp 2048–2051
57. Snidaro L, Micheloni C, Chiavedale C (2005) Video security for ambient intelligence. IEEE Trans Syst Man Cybern Part A Syst Humans 35(1):133–144
58. Bagaa M, Taleb T, Ksentini A (2016) Efficient tracking area management framework for 5G networks. IEEE Trans Wirel Commun 15(6):4117–4131
59. Pang Y, Yuan Y, Li X, Pan J (2011) Efficient hog human detection. Sig Process 91(4):773–781
60. Lowe DG (1999) Object recognition from local scale-invariant features. In: Proceedings of the Seventh IEEE International Conference on Computer Vision, vol 2. IEEE, pp 1150–1157
61. Choi T-W, Kim D-H, Kim K-H (2016) Human detection in top-view depth image. Contemp Eng Sci 9(11):547–552
62. Ahmed I, Ahmad M, Adnan A, Ahmad A, Khan M (2019) Person detector for different overhead views using machine learning. Int J Mach Learn Cybern 10(10):2657–2668. https://doi.org/10.1007/s13042-019-00950-5
63. Ullah K, Ahmed I, Ahmad M, Khan I (2019) Comparison of person tracking algorithms using overhead view implemented in opencv. In: 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON). IEEE, pp 284–289
64. Ahmed I, Ahmad M, Nawaz M, Haseeb K, Khan S, Jeon G (2019) Efficient topview person detector using point based transformation and lookup table. Comput Commun 147:188–197
65. Du D, Qi Y, Yu H, Yang Y, Duan K, Li G, Zhang W, Huang Q, Tian Q (2018) The unmanned aerial vehicle benchmark: object detection and tracking. In: European Conference on Computer Vision. Springer, New York, pp 375–391
66. Zhu P, Wen L, Du D, Bian X, Ling H, Hu Q, Wu H, Nie Q, Cheng H, Liu C et al (2018) Visdrone-vdt2018: The vision meets drone video detection and tracking challenge results. In: Proceedings of the European Conference on Computer Vision (ECCV)
67. Qi Y, Zhang S, Zhang W, Su L, Huang Q, Yang M-H (2019) Learning attribute-specific representations for visual tracking. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 33, no 1, pp 8835–8842
68. Ahmad M, Ahmed I, Adnan A (2019) Overhead view person detection using yolo. In: IEEE 10th Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON), pp 0627–0633
69. Ahmad M, Ahmed I, Ullah K, Ahmad M (2019) A deep neural network approach for top view people detection and counting. In: IEEE 10th Annual Ubiquitous Computing, pp 1082–1088
70. Ahmed I, Din S, Jeon G, Piccialli F (2019) Exploring deep learning models for overhead view multiple object detection. IEEE Internet Things J 7(7):5737–5744
71. Ahmad M, Ahmed I, Khan FA, Qayum F, Aljuaid H (2020) Convolutional neural network-based person tracking using overhead views. Int J Distrib Sens Netw 16(6):1550147720934738
72. Ahmed I, Ahmad M, Khan FA, Asif M (2020) Comparison of deep-learning-based segmentation models: Using top view person images. IEEE Access 8:136361–136373
73. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds) Advances in neural information processing systems. Curran Associates, Inc., Red Hook, pp 1097–1105. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf
74. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
75. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9
76. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 580–587
77. Girshick R (2015) Fast R-CNN. In: Proceedings of the IEEE international conference on computer vision, pp 1440–1448
78. Ren S, He K, Girshick R, Sun J (2015) Faster R-CNN: towards real-time object detection with region proposal networks. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (eds) Advances in neural information processing systems. Curran Associates, Inc., Red Hook, pp 91–99
79. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: common objects in context. In: European conference on computer vision. Springer, New York, pp 740–755 (**ISBN: 978-3-319-10602-1**)
80. West J, Ventura D, Warnick S (2007) Spring research presentation: a theoretical foundation for inductive transfer, vol 1, no 8. Brigham Young University, College of Physical and Mathematical Sciences
81. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 779–788. arXiv:1506.02640
82. Zhang X, Yang W, Tang X, Liu J (2018) A fast learning method for accurate and robust lane detection using two-stage feature extraction with yolo v3. Sensors 18(12):4308
83. Bernardin K, Stiefelhagen R (2008) Evaluating multiple object tracking performance: the clear mot metrics. EURASIP J Image Video Process 2008:1–10
84. Milan A, Leal-Taixé L, Reid L, Roth S, Schindler K (2016) Mot16: a benchmark for multi-object tracking. arXiv arXiv:1603.00831