



Automated Physical Distance Estimation and Crowd Monitoring Through Surveillance Video

Masum Shah Junayed^{1,2} · Md Baharul Islam^{1,2,3}

Received: 15 March 2022 / Accepted: 24 October 2022 / Published online: 24 November 2022
© The Author(s), under exclusive licence to Springer Nature Singapore Pte Ltd 2022

Abstract

The contagious Corona Virus (COVID-19) transmission can be reduced by following and maintaining physical distancing (also known as COVID-19 social distance). The World Health Organisation (WHO) recommends preventing COVID-19 from spreading in public areas. On the other hand, people may not be maintaining the required 2-m physical distance as a mandated safety precaution in shopping malls and public places. The spread of the fatal disease may be slowed by an active monitoring system suitable for identifying distances between people and alerting them. This paper introduced a deep learning-based system for automatically detecting physical distance using video from security cameras. The proposed system introduced the TH-YOLOv5 for object detection and classification and Deepsort for tracking the detected people using bounding boxes from the video. TH-YOLOv5 included another prediction head to identify objects of varying sizes. The original prediction heads are then replaced with Transformer Heads (TH) to investigate the prediction capability of the self-attention mechanism. Then, we include the convolutional block attention model (CBAM) to identify attention areas in settings with dense objects. Pairwise L2 vectorized normalization was utilized to generate a three-dimensional feature space for tracking physical distances and the violation index, determining the number of individuals who follow the distance rules. We use the MS COCO and HumanCrowd, CityPersons, and Oxford Town Centre (OTC) data sets for training and testing. Experimental results demonstrate that the proposed system obtained a weighted mAP score of 89.5% and an FPS score of 29; both are computationally comparable.

Keywords COVID-19 social distancing · Human detection and tracking · Distance measurement · Crowd monitoring · Video surveillance

Introduction

Corona Virus (COVID-19) is a highly contagious worldwide virus that has wreaked havoc globally. It has created distance among people. The Corona Pandemic affected about 220 nations and regions worldwide, with roughly 197,201,247 confirmed cases since July 2021 [1]. Numerous COVID-19 virus variants have already been identified worldwide. This virus is continuously evolving due to mutations. Recently, scientists discovered a novel variation in India, dubbed the Delta variant. The Delta strain spreads at a rate that is several times that of other strains [2]. It is dominating the world health sector.

To prevent the massive spread of COVID-19, the WHO establishes specific fundamental guidelines for humans, such as maintaining physical distance from others, wearing a face mask, washing hands for 20 s, wearing PPE, and staying at home. However, it is difficult for the government to control

This article is part of the topical collection “Advances in Applied Image Processing and Pattern Recognition” guest edited by K C Santosh.

✉ Masum Shah Junayed
masumshahjunayed@gmail.com

Md Baharul Islam
bislam.eng@gmail.com

- ¹ Department of Computer Engineering, Bahcesehir University, Istanbul 34349, Turkey
- ² Department of Computer Science and Engineering, Daffodil International University, Dhaka 1207, Bangladesh
- ³ College of Data Science and Engineering, American University of Malta, Bomla 1013, Malta

individuals in public areas. As a result, governments in many countries altered their policies to minimize the number of people in public places, such as closing all educational institutions, limiting person attendance in the workplace, and maintaining a distance of at least 2 m for garment workers wearing a face mask. Furthermore, several software firms established their internet offices from home. The physical distance between people contributes to the reduction of COVID-19 virus transmission. Numerous institutions initiate a program to remove individuals from densely populated regions to reduce the COVID-19 spread. For instance, the government puts police officers in public areas to ensure public safety, security, and physical distance. The traditional approach (check individuals by security personnel) are tedious, time-consuming, and imprecise. The primary purpose of doing this study is to determine the distance between individuals using a real-time video, which also helps to protect the particular place from the spreading of COVID-19. Our system effectively identifies physical spaces and improves the safety from COVID-19 in specific regions, such as constructions and garments areas.

The technology is expected to use the latest machine learning technology, and surveillance cameras in the building to identify whether or not individuals are keeping a safe physical distance from each other based on real-time video feeds [3]. This technology is also said to connect with security cameras at various businesses to prevent employees from working too closely together. Three demonstration steps were provided, each representing a stage in the process of calibration, detection, and measurement.

The development of a wide variety of numerical techniques and models has allowed us to evaluate the evolution of pandemic processes throughout the last 2 years due to research on the assessment of physical distance in COVID-19 [4–8]. They spoke about how people’s physical distance is affected by their social environment, and they brought up COVID-19’s worry. Recently, some commercialized computer vision-based systems [9–11] for monitoring physical distance have been established. These methods are intriguing; however, there is no statistical analysis included in the findings. In addition, there is little discussion of implementation. While the conversations are enlightening, they do not provide concrete results for measuring physical distance, leave the issue open, and may be considered undesirable by others.

Nevertheless, none of these approaches provide in-depth explanations of their methods, performance benchmarks, or logic for their detection algorithm selection. Some work that one must use a theoretical approach for physical distance, but leave out specifics of what steps can be taken in real life. In contrast, we have suggested a real-time automated surveillance system for detecting and monitoring people and measuring inter-distance between humans, assessing risk in

real-time by warnings in the form of red text (‘Unsafe’) and counting them using the bounding boxes. Furthermore, our technology is crucial for observing the physical distance between peoples in density areas and controls entry to a particular location. Listed below are the most important contributions of this paper:

- We introduced a deep learning-based automated system for monitoring and detecting people to reduce coronavirus expansion and economic costs.
- A transformer-based TH-Yolov5 is used to detect and classify Pedestrians, while Deepsort [12] is used to track people in this system.
- We used Transformer Heads (TH) to detect pedestrians in high-density environments. In addition, The CBAM was included in YOLOv5 to assist the network in locating areas of interest in images with extensive region coverage.
- We utilized pairwise L2 vectorized normalization, which uses the centroid coordinates and dimensions of the bounding box to create 3D feature space. We calculated how many people are not following the physical distance regulation using the violation index.

The following sections comprise the rest of this paper: Section “[Related Works](#)” summarizes the relevant literature; Sect. “[Proposed Method](#)” elaborates on the proposed approach; Section “[Experiments](#)” explains the data sets and training details; Section “[Results and Discussions](#)” elaborates the specifics of our results and discussions. Finally, we conclude this article in Section “[Conclusion](#)”.

Related Works

COVID-19 affects individuals differently, but it mainly spreads via droplet contact, physical touch, and airborne transmission. physical distance may play an essential role in reducing the spread of COVID-19 [13]. As a result, everyone should be careful and observe the norms of physical distancing, such as keeping a set space (100 cm) between themselves and others. That is why it is also known as “Physical Distancing”. Several techniques for detecting objects from movies and pictures have previously been suggested for a variety of applications [14]. This section has been discussed comprehensively in the literature review.

Bouhleh et al. [15] developed a video sequence technique that combines two approaches, the macroscopic and microscopic methods, to calculate the real-time distance between individuals. Based on the techniques, they utilized two kinds of data sets, Mayenberg and Mliki. They used a three-level categorization method to improve performance substantially; however, it is not suited for real-time use.

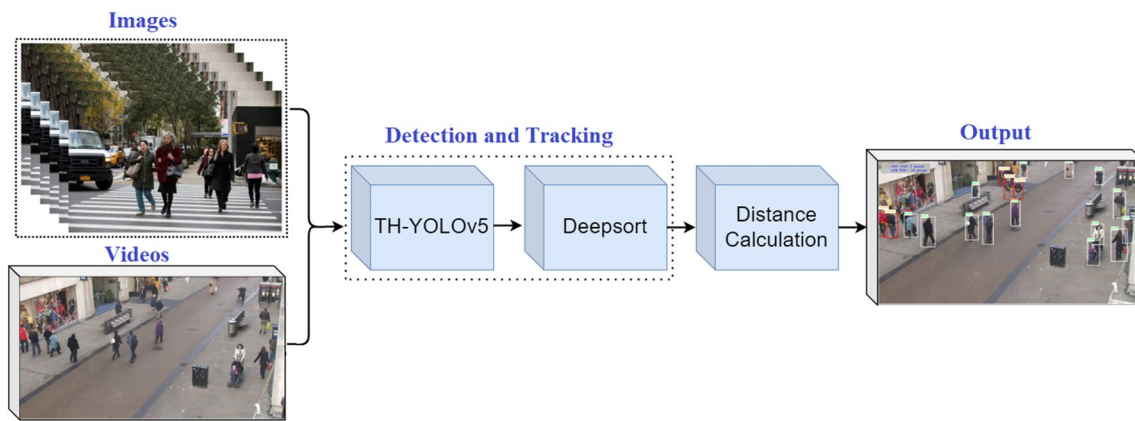


Fig. 1 Block diagram of the proposed physical distancing system. The proposed TH-YOLOv5 used for pedestrian Detection and classification, deepsort and pairwise L2 norm are used to tracking and physical distancing

In the COVID-19 scenario, Razavi et al. [16] devised an automated method to monitor construction workers to guarantee their safety. When the employees are on duty, the system detects the face mask and the physical distance between them. They utilized Faster R-CNN Inception ResNet V2 for image detection to improve the system's accuracy. However, as the quantity of training data decreases, the accuracy decreases. Furthermore, the system is unable to detect the mask when the employees swivel their heads. They used Faster R-CNN Inception V2 to detect the distance between people. They calculated the actual distance traveled by the workers from the image in which the established method did not work effectively, because their performance is not good.

Rahim et al. [17] proposed a technique for measuring physical distance in low-light environments. COCO location measurements were used to assess the trained model's performance. After utilizing the YOLO-v4 model for real-time object recognition and physical distance measurement, but it is not real time. Two-stage locators achieve better restriction and item recognition accuracy, while one-stage locators achieve faster deduction speed [16]. The technology only works in settings with a fixed physical distance and two target items. Some prototypes that use machine learning and sensing technologies for physical distance tracking have been proposed. Landing AI [3] suggested a physical distance estimator that uses a security camera to identify individuals whose physical distance is less than the acceptable value. In a manufacturing facility, another system [18] was used to control and tracking labor movements and deliver real-time audio warnings. Along with security cameras, systems based on LiDAR and stereo cameras [19] were presented, demonstrating that other kinds of sensors than monitoring cameras can also be beneficial.

Using YOLOv3 and Deepsort, a new approach was presented for locating and tracking individuals [20] that involves monitoring the user's social distance. The

procedure also measures the extent to which other non-social-distancing activities occur and then calculates an index of non-social-distancing behaviors. This method seems unique. However, it lacks any statistical analysis.

Rezaei et al. [6] introduced the DeepSOCIAL, which is the DNN-based model. It uses cutting-edge deep learning methods to identify, track, and measure social distances. Furthermore, it detected social distance and used dynamic risk assessment. While this paper did not address pure violation detection, it did discover a method to help reduce congestion.

While the methods described above are intriguing, collecting data and issuing invasive warnings may be considered undesirable by some individuals. On the contrary, we present an automated system capable of real-time human identification, tracking, and physical distance measurement and indicating whether a person is safe or unsafe.

Proposed Method

This section describes our approach system for physical distancing monitoring consisting of three stages: detection and tracking of persons, inter-distance calculation, and zone-based infection risk evaluations. The system is designed to operate with and is suitable for all types of CCTV security cameras, independent of video quality and real-time depiction of identified people and their distance. Figure 1 shows the block diagram of the proposed system. The system is divided into three parts: detection, tracking, and distance estimation. The subsections show the full process of our suggested system.

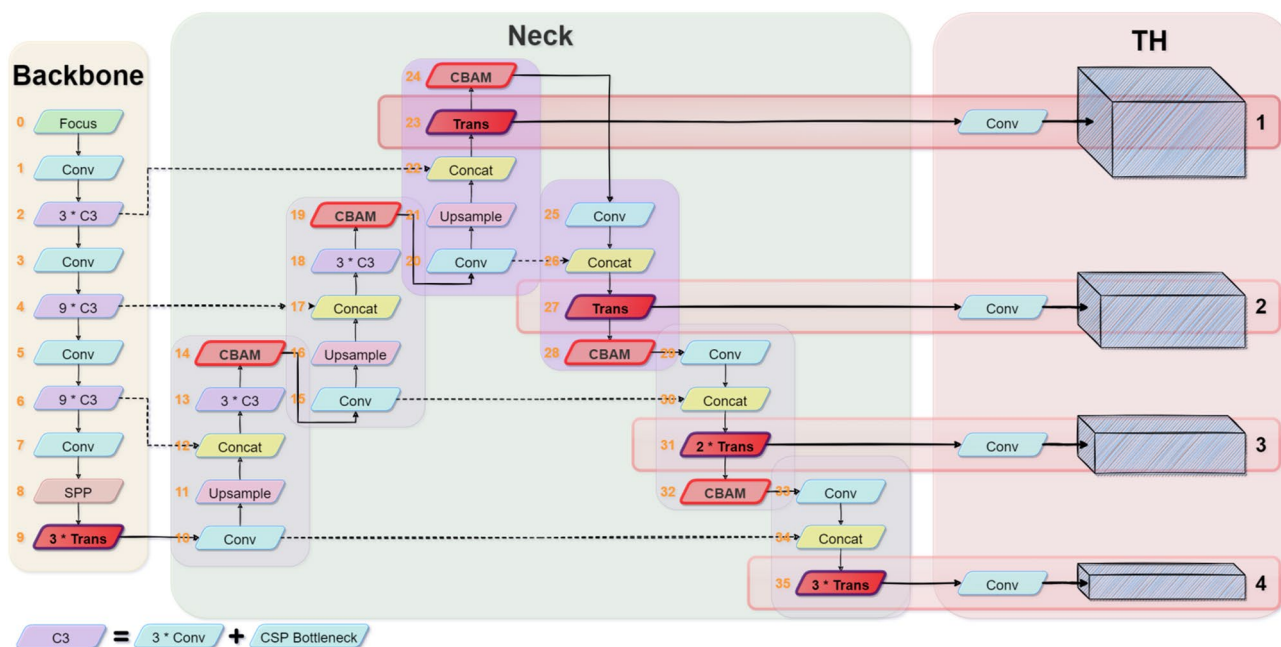


Fig. 2 Architecture of the proposed TH-Yolo-v5 for pedestrian detection. It has three main modules including backbone, neck and head

Pedestrian Detection

A real-time human detector is introduced for the complex environment that contains various objects to identify people correctly. This detector has a feature extractor and classification module, which is based on Yolo-v5 [21]. A fundamental strategy of improving the exactness feature extraction of CNN-based object detectors [22] is to broaden the receptive field and increase the system’s complexity employing additional layers whereas identifying tiny objects more effectively. Instead, for easier training, we utilized a skip-connections method. Consequently, to decrease the size of the parameter, a modified CSPDarknet53 with SPP layers is utilized as a feature extractor backbone. Then, PANet [23] serving as the neck, and YOLO serving as the detecting head [24]. A collection of freebies and specials [25] is supplied to optimize the whole architecture. It reprocesses and logically employs the feature maps collected by Backbone at various phases. A neck is usually made up of multiple bottom-up and top-down routes. The neck is an essential connection in the target detection architecture. After that, a transformer head (TH) used as classification network that is intended to extract features maps from the backbone to detect the location and classification of the person. The transformer encoder is utilized in Head part. Figure 2 depicts the TH-YOLOv5 architecture.

Transformer Head (TH). We analyze the CrowdHuman [26] and CityPersons [27] data sets discover a large number of minor occurrences; therefore, we add another prediction head for detecting small items. When combined

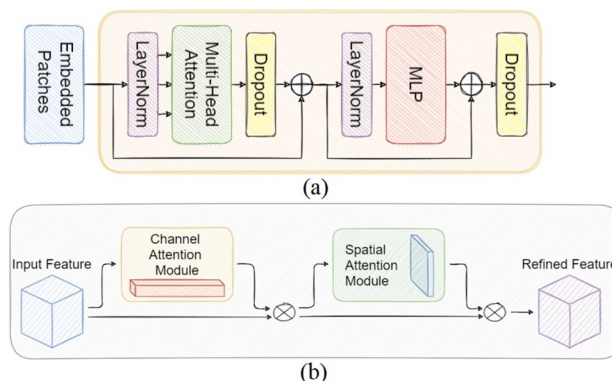


Fig. 3 Architecture of the transformer encoder module (a) and CBAM module (b)

with the other three prediction heads, our four-head structure can reduce the detrimental effect produced by violent object scale variation in the predictions. The prediction head (head No.1) that we add is created using a low-level, high-resolution feature map that is more sensitive to tiny objects, as illustrated in Fig. 2. The performance of tiny object detection improves significantly after adding a detection head, even though the compute and memory costs rise.

Transformer Encoder. We modified the original version of YOLOv5 to include transformer encoder blocks, which were inspired by the vision transformer [28]. We replaced several convolutional blocks and CSP bottleneck blocks with transformer encoder blocks. Figure 3a depicts the structure. Comparing the transformer encoder block with

CSPDarknet53 to the original bottleneck block in CSPDarknet53, it can collect more global information and a great deal of contextual information. There are two sub-layers in each transformer encoder. It is divided into two sub-layers: the first is a multi-head attention layer (MLP), and the second MLP is fully connected. Between each sublayer, residual connections are employed. Transformer encoder blocks improve the ability to record a variety of local objects. It performs better on occluded persons with a high density on the CityPersons data set. We used multi-head attention, a unit for attention mechanisms that simultaneously processes an attention mechanism multiple times. In addition, it may be used for research on the feature representation possibilities. The outputs of the particular attention are then concatenated and converted linearly into the predicted dimension. Intuitively, having many attention heads enables one to focus on various sections of the sequence in different ways. It allows the model to continuously attend to input originating from various representation subspaces located at multiple places. A single attention head is required for averaging; thus, this is inhibited. The multi-head (MH) is followed as follows:

$$\text{MH}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_n)P^O \quad (1)$$

$$\text{head}_i = \text{Attention}(QP_i^Q, KP_i^K, VP_i^V) \quad (2)$$

where Q , K , and V depict the query, key and value, respectively. In addition, P_i^Q , P_i^K , and P_i^V are represent the matrices of the parameter, respectively. In addition, $h = 2$ parallel attention layers or heads are used to the minimizing dimension.

We used transformer encoder blocks in the head section of the backbone for human prediction. The Transformer Head (TH) is used based on YOLOv5, because the feature maps have a low resolution after the end of this backbone network. TH can reduce computational and memory costs when used on low-resolution feature maps. Furthermore, as we increase the resolution of the input images, we can eliminate certain TH blocks from the early layers to enable the training process.

Convolution Block Attention Module (CBAM). CBAM [29] is a simple yet very effective attention module. It is a lightweight module inserted into most well-known CNN architectures and trained in an end-to-end technique. CBAM performs adaptive feature refinement by progressively inferring the attention map along two different channel dimensions and then multiplying the attention map with the input feature map. CBAM uses a feature map as an input for adaptive feature improvement and then progressively implies the attention map along two distinct dimensions: channel and spatial. It then combines the attention map with the feature maps. Figure 3b illustrates the structure of the CBAM module. Large coverage regions in CCTV collected photos usually feature perplexing geographical characteristics. The

CBAM is used to extract the attention area, which can benefit TH-YOLOv5 in avoiding misleading input and focusing on valuable target items.

Human Tracking

Deepsort [12] is used to track people in any video in human tracking. It is created using discovered from identified humans in pictures pattern, which subsequently is coupled with temporal data to forecast the subjects' trajectories. It maps unique IDs to keep track of each item under investigation for statistical analysis. Deepsort can also be used to deal with occlusion, numerous perspectives, and annotation of training data. In Deepsort, Kalman filter and Hungarian algorithm have been commonly employed for accurate tracking. For improved association, the Kalman filter is employed recursively, and it can forecast future locations based on present positions [30]. We subsequently utilize this time information to assess the severity of physical distance breaches and the presence of high-risk zones on the scene. The status of each individual in a frame is represented by the following this equation:

$$F = [x, y, a, b, xt, yt, at]^T \quad (3)$$

where (x, y) indicates the target bounding box's horizontal and vertical positions; a signifies the scale (area); and b specifies the bounding box's aspect ratio. xt , yt , at are the anticipated values for the horizontal area, vertical area, and bounding box centroid, respectively, as predicted by the Kalman filter.

We performed the following matrix D_t , which contains the position of the n identified persons in the image carrier grid: After completing the detection and tracking procedure (P^t), for each input frame $w * h$ at time t , we define the following matrix:

$$D_t = \{P_{x_n, y_n}^t \mid x_n \in w, y_n \in h\} \quad (4)$$

Distance Measurement

Researchers have developed various 2D and 3D depth estimation techniques [31, 32]. We calculated the distance between identified tracked individuals on each image and video.

The deepsort model generates a collection of bounding boxes and an ID for each person detected in the previous phase. When working with bounding boxes (such as rectangular boxes), coordinates (x, y) in the 3D (x, y, d) feature space. However, to the picture that was obtained from the camera, the reduced 2D space of (x, y) correlates to two parameters (x, y) alone, and depth (z) is not accessible. To better visualize the 3D shape of each bounding box, imagine every point in space equates to three values (x, y, z) . Using this Eq. 5, the 2D pixel

coordinates (x, y) as an input, the world coordinate points are then mapped to the points on the screen (X_w, Y_w, Z_w) [33].

$$[x, y, 1]^T = KRT[X_w, Y_w, Z_w, 1]^T \quad (5)$$

Here, K , R , and T represent the rotation, translation, and intrinsic matrix, respectively. This feature space indicate the coordinates of the centroid, and the value of d describes the depth of each object [34]. We computed the following equation 6 to estimate the depth between camera and objects, which can be acquired by studying the form of the image [35].

$$d = \frac{2\pi \times 180}{(w \times h \times 360) \times 1000 + 3} \quad (6)$$

where w denotes the bounding box's width, and h denotes the bounding box's height. The pairwise L2 normalization is calculated for the collection of bounding boxes as provided by the following Eq. 7.

$$D = \sqrt{\sum_{i=1}^n (q_i - p_i)^2} \quad (7)$$

where we define this equation $n=3$. Here, D represent the distance and q_i and p_i represents pixels (range 90 to 170) between two humans. After locating the individual's neighbors using the L2 norm, we allocate them based on their proximity sensitivity. The proximity threshold is constantly updated with a large number of tests using a set of numbers between 90 to 170 pixels wide, depending on the person's position in a particular frame. To use the proximity property, every person in the system must be given at least one neighbor or many additional neighbors to create a group in distinct color-coding. The creation of groups implies the breach of the physical distance practice, which is measured using the following equation:

$$v_i = \frac{n_p}{n_g} \quad (8)$$

where v_i represents the index of violation. To calculate the number of groups or clusters present in the video, the number of separate groups or clusters detected (n_g) and n_p defines the total number of individuals close to those groups or clusters nearby.

Experiments

Data sets and Evaluation Matrix

Table 1 Specifics of the data sets that were utilized in our experiments. The number of images assigned to each data set, train/test, and pedestrian is shown here

Data sets	Images	Persons	Training	Test
MS-COCO [36]	64,115	257,215	64,115	–
CrowdHuman [26]	15,000	339,565	–	15000
CityPersons [27]	2975	257,252	–	2975
OTC [37]	CCTV videos	2200	–	CCTV videos

The pedestrian recognition, tracking, and distance estimation problems center on the full data set. It gives the same database for researchers to use when comparing the effectiveness of various algorithms. It serves as a source of data for researchers, so that they may carry out experimental experiments. When evaluating the quality of a data set, it is essential to consider both the quantity of data and the accuracy of the labeled information. The resilience of the detector is, to some part, determined by the depth of the data set being analyzed. When compared to tasks involving the recognition of broad objects, pedestrian detection possesses distinct properties. The proposed system is trained on object detection MS COCO data set [36] that has 80 classes and 123k images. The bounding box labels on each image were additionally annotated with the matching coordinates. The CrowdHuman [26] and CityPersons [27] data sets are also used to person detection to assess our system. Both data sets include annotations for two types of bounding boxes: visible human area and human full-body bounding boxes. The CrowdHuman data set is far more challenging to work with than CityPersons, since it has more instances per image, and those examples are often strongly overlapped. Table 1 illustrates an overview of the information included in the data set. To compare and evaluate the proposed system, we used another data set, the Oxford town Centre (OTC) data set [37], an unknown and complex data set with a high frequency of object recognition overlapping and overcrowded zones. The collection also included a wide range of human clothing and looks in the public location real world. FPS, mAP, and total loss in identifying the individual, as shown in Fig. 5, are constantly measured throughout the validation period.

Traning Details

All training and testing performed on the same PC, which has a Windows 10 operating system, an Intel Core i9-10850K CPU running at 3.6 GHz, 64 GB of RAM, and an NVIDIA GeForce RTX 2070 super GPU with 8GB video RAM. We utilized PyTorch, a CUDA environment, and a vscode editor to implement it. Following training using TH-YOLOv5, we evaluated the data sets and assessed the

Table 2 Pedestrian Detection performance of three benchmark data sets with two well-known backbones in terms of mAP, NOI, Total loss and FPS

Data set	Backbone	mAP	NOI	Total loss	FPS
CrowdHuman [26]	CSPResNext50	0.913	16246	0.87	28
	CSPDarkNet53	0.921	16049	1.14	28
CityPersons [27]	CSPResNext50	0.937	12681	1.64	29
	CSPDarkNet53	0.928	12523	0.59	27
OTC [37]	CSPResNext50	0.887	13260	0.95	30
	CSPDarkNet53	0.895	13847	1.08	29

The bold indicates the best performance

outcomes by exhibiting the failure occurrences. We determined that TH-YOLOv5 has excellent localization capability but limited classification capability.

The transfer learning concept was used to train the TH-YOLOv5-based model, then fine-tuned and optimized before being used to train the proposed model on the MS COCO data set. We utilized SGD with warm restarts to alter the learning rate throughout the training phase. It aided in breaking out of local minima in the solution space and saving training time. The technique started with a high learning rate, slowed it down midway, and then decreased the learning rate for each batch with a slight downward slope. It has the effect of jumpstarting the solution space out of local minima, saving the training time. A high learning rate was used in the beginning stage and slowed down midway. Progressively, the learning rate was decreased for each batch, with a minimal drop in speed.

Results and Discussion

Performance of Pedestrian Detection

Table 2 demonstrates the performance of the Pedestrian Detection on three popular benchmark data sets using our proposed model, TH-YOLOv5. For every data set, we have employed two backbones in TH-YOLOv5, including CSPResNext50 and CSPDarkNet53. From Table 2, it can be observed that the higher mAP is obtained in every data set through our proposed TH-YOLOv5, mainly when CSPResNext50 and CSPDarkNet53 are utilized in CrowdHuman [26], the mAP is 91.3% and 92.1% with a minimum loss of 0.87 and 1.14, around 16k iterations and 28 FPS, respectively. Similarly, in CityPersons, and OTC, the TH-YOLOv5 shows outstanding accuracy, i.e., 93.7%, 92.8%, 88.7%, and 89.5% in terms of using CSPResNext50 and CSPDarkNet53 backbones. With the minimum loss, the

number of iterations is also less (about 12.5k and 13k) in CityPersons [27] and OTC [37] data sets.

Figure 4 displays the visual findings that were achieved by the suggested TH-YOLOv5. The optical pedestrian detection of the CrowdHuman [26], CityPersons [27], and OTC [37] data sets is shown in the first, second, and third rows, respectively. In the first and second rows, we can see that our suggested model can adequately detect pedestrians. The suggested TH-YOLOv5 can identify tiny things (humans/pedestrians) from the perspective of a surveillance camera in the third row.

Performance Analysis

We calculated the evaluation matrices of mean average error (MAE) and average closest physical distance, d_{avg} for the overall frame for measuring physical distance. We calculated the $d_{avg} = \frac{1}{N} \sum_{i=1}^m d_i^{min}$, where $d_i^{min} = \min(d_{i,j})$, $\forall j \neq i \in \{1, 2, \dots, n\}$ is the closest physical distance between number of humans and the MAE of the social distancing violation ratio $r_v = \frac{v}{n}$, where v is the number of humans who break the physical distance and n is total number of people.

The performance of physical distance (also known as COVID-19 social distance) is shown in Table 3. It calculates the amount of time it takes to identify breaches of physical distance. The proposed TH-YOLOv5 with CSPResNext50 and CSPDarknet53 backbones is showed the performance of measuring physical distance and violation of physical distance on the OTC data set. In the CSPResNext50 backbone, the proposed TH-YOLOv5 has achieved 1.514 d_{avg} and 0.167 r_v of MAE. Our proposed method outperforms the CSPDarknet53 backbone, obtaining an MAE of 0.559 d_{avg} and 0.114 r_v on the OTC data set. Figure 5 shows the lowest and average closest distances. The blue line shows the average number of individuals who are near each other, while the orange line represents the lowest number of people in this data set.

Figure 6 shows how the frequency of social distance violations correlates with the social density ($human/m^2$) in 2D histograms on the OTC data set. A violation rises with a rise in social density. It can be seen in the graph, where the two variables (social density and violation) are linearly related to one another. In addition, it is possible to use the suggested linear regression.

Figure 7 depicts the detection and physical distancing outcomes of the proposed approach. The distance between people is shown in this diagram. It shows the safe and count number with a white bounding box if the distance between two people is more than 100 cm. On the other hand, the bounding box was red with the 'unsafe' word and contained a count of dangerous people from the input video.

Figure 8 depicts the analysis of overlapping activation map. According to this diagram, the level of violation

Fig. 4 Visual representation of pedestrians detection. The detection of results obtained from the CrowdHuman [26], CityPersons [27], and OTC [37] data sets are shown in the figure's first, second, and third rows, respectively

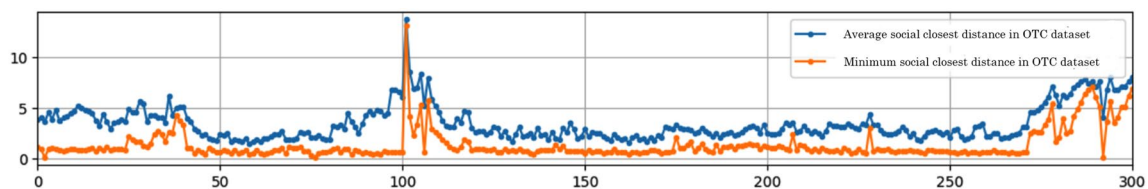


Fig. 5 Minimum vs. average physical closest distance in Oxford Town Center data set

Table 3 Performance of physical distance

Methods	MAE of d_{avg} (meter)	MAE of r_v (count)
TH-YOLOv5 (CSPResNext50)	1.514	0.167
TH-YOLOv5 (CSPDarknet53)	0.559	0.114

The bold indicates the best performance

increases as the blue color moves closer and closer to the red color. The first, second and third rows are represent the flow of pedestrian, violation heatmap and risky areas, respectively. Figure 8a, b represents two examples of situations in which two persons remained stationary in two specific spots of the grid for an extended length of time; as

a result, the heat map became reddish for those locations. Because there was more foot activity on the sidewalks than in the center of the roadway, the heat maps for those areas revealed much higher temperatures. In general, the more red grids that are present may suggest locations that are possibly more dangerous. It would be more advantageous to analyze the density and position of the individuals who significantly breached the physical distancing measures, in addition to people's natural movement and tracking data. In addition, for a better 3D portrayal of safe and dangerous zones, (a) through (c) correlate to people's flow and violation heatmaps.

The visual investigation of activation heatmap are shown in Fig. 9. In this figure, (a), (b), (c), and (d) demonstrate the human detection and tracking (the green and yellow bounding boxes represent the safe and dangerous region), the

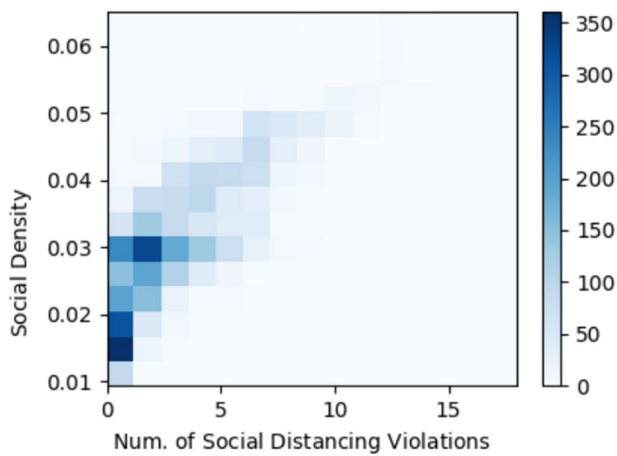


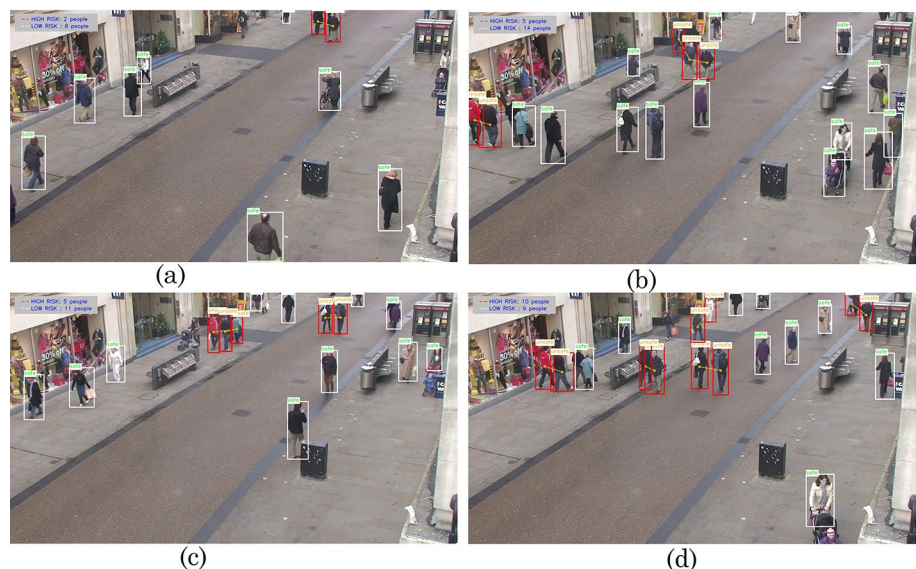
Fig. 6 2D histograms that visualize the frequency of social distance violations vs. social density. The observation on the histograms correlates with a positive correlation

mosaic heatmap of violation, the crowd violation heatmap, and the live people status data, respectively. The crowd map (Fig. 9c) displays the map showing infractions and risks, where we applied trends for rising and falling decay levels. In addition, the visual graph (Fig. 9d) offers basic information about the population inside each frame, the number of persons who fail to observe the physical distance, and the number of social distance violations without including coupled groupings.

Comparative Analysis and Discussion

Table 4 shows the performance of human detection with five well-known object detectors and the proposed TH-YOLOv5. The TH-YOLOv5 is evaluated in the proposed system using

Fig. 7 Output of the proposed physical distancing system. Here, (a)–(d) represents the low risk, lower medium risk, medium risk and high risk



two alternative backbones: CSPResNext50 and CSPDarknet53. The faster RCNN [38], YOLO-v3 [38], YOLO-v4 [25], EfficientDet [40] and the YOLO-v5 [21] models are utilized to compare with the proposed TH-YOLOv5 for human detection and tracking. The Faster RCNN attained an accuracy of 0.813 mAP and a frame rate (FPS) of 27. The YOLO-v3 achieved an mAP of 0.839 with a frame rate of 26 FPS. It outperforms faster RCNN, while falls shorter FPS. This table also illustrates the performance of the proposed YOLO-v4, and YOLO-v5 with the two backbones and EfficientDet. The mAP, iteration number, training loss (TL), and FPS are compared. The results of the TH-YOLOv5 experiment with and without the transformer head (TH) are shown here. Both of these studies use the CSPResNext50 and CSPDarknet53 backbones, respectively. In the model provided, it can be seen that the values of mAP obtained with TH were the greatest (0.895 with CSPDarknet53). The suggested model with TH reached its maximum values while suffering only a tiny bit of a drop in terms of the number of iterations (NOI) (13847 and 1.08 in CSPDarknet53, respectively). In addition, The CSPResNext50 achieves the highest FPS of 30.

To illustrate the efficiency of the suggested approach, we compared our findings to those of previous techniques in Table 5. Rezaei et al. [6], and Yang et al. [7] presented the techniques and obtained fps of 24.1 and 25 with 512×512 and 1920×1080 input sizes, indicating that this method works well when the input size is big. In the suggested system, our input size is just 640×640 . However, our FPS is 30. Therefore, the proposed approach achieves high FPS on the same OTC data set and received better than the state-of-the-arts.

Table 6 shows the comparison between our proposed system and the recent work [42]. We can observe from this table that our model (TH-YOLOv5 with CSPDarkNet53) consumes less frames per second (300 vs. 33), but our average

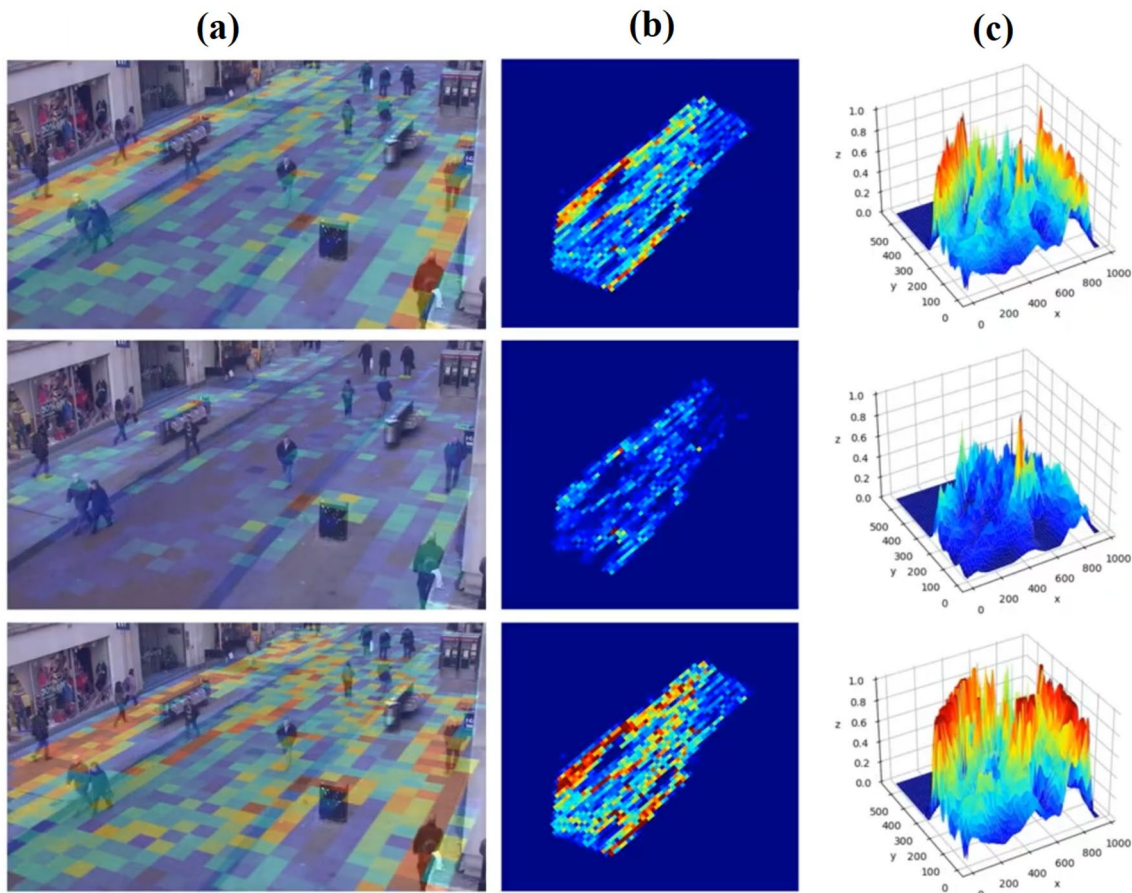


Fig. 8 Analysis of the overlapping activation map using the TH-YOLOv5 that was suggested. Here, the first, second, and third rows show the flow of people, violations, and dangerous areas. In addition,

(a)–(c) illustrate the mosaic heatmap, the violation heatmap, and the 3D tracking + violation heatmap, respectively

Fig. 9 Visual investigation of the activation heatmap is shown. In this case, (a)–(d) refer to the pedestrian detection and tracking, mosaic heatmap, crowd heatmap, and people status statistics, respectively

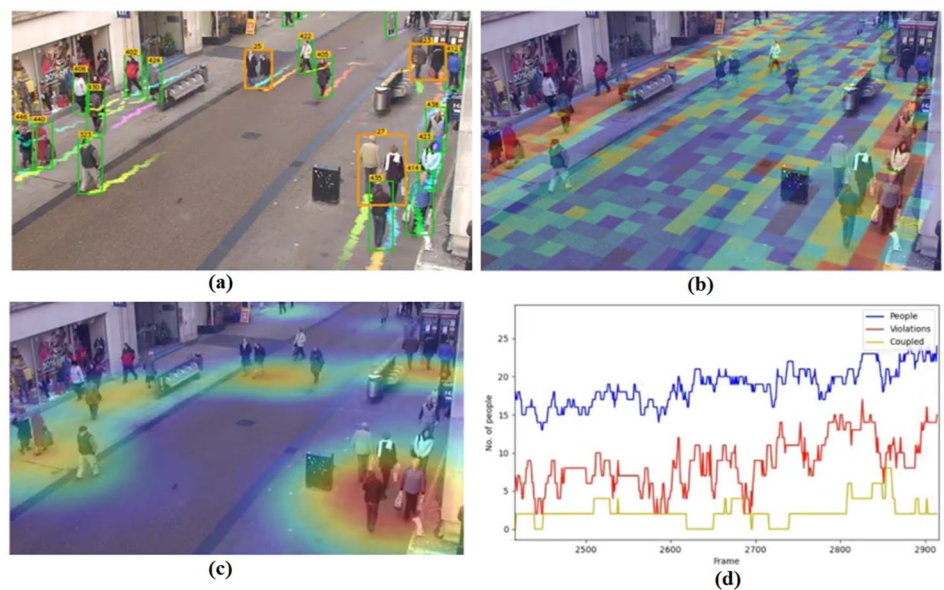


Table 4 Five well-known object detectors and the proposed TH-YOLOv5 with two backbones are compared in real-time human detection

Method	Backbone	mAP	NOI	Total loss	FPS
Faster RCNN [38]	–	0.813	12538	1.98	27
YOLO-v3 [39]	Darknet53	0.839	12135	1.85	26
YOLO-v4 [25]	CSPResNext50	0.855	12847	0.96	29
	CSPDarknet53	0.868	13455	1.81	28
EfficientDet [40]	Darknet53	0.862	13169	1.77	29
YOLO-v5 [21]	CSPResNext50	0.863	12988	0.94	28
	CSPDarknet53	0.879	13622	1.52	28
TH-YOLOv5 (Without TH)	CSPResNext50	0.865	13003	0.94	28
	CSPDarknet53	0.878	13781	1.57	27
TH-YOLOv5	CSPResNext50	0.887	13260	0.95	30
	CSPDarknet53	0.895	13847	1.08	29

The bold indicates the best performance

Table 5 Comparison of the proposed system with the state-of-the-art methods to perform real-time physical distancing measures at frames per second (FPS)

Approaches	Data set	Input size	FPS
Pouw et al. [5]	OTC	–	10
Punn et al. [20]	OTC	–	23
Rezaei et al. [6]	OTC	512 × 512	24.1
Yang et al. [41]	OTC	1920 × 1080	25
Saponara et al. [7]	OTC	–	27
Yadav et al. [42]	OTC	–	27
TH-YOLOv5	OTC	640 × 640	30

The bold indicates the best performance

Table 6 Comparison of the proposed TH-YOLOv5 with the recent system in terms of AP_{50} , input size and FPS

Methods	AP_{50}	Input size	FPS
Yadav et al. [42] (2022)	40.2	600 × 600	300
TH-YOLOv5	88.1	600 × 600	33

The bold indicates the best performance

precision is much better (40.2 vs. 88.1) while having the same input size of 600 × 600.

Conclusion

This article introduced a deep learning-based automated system that works on real-time physical distance monitoring, detecting, and tracking people using bounding boxes. We presented the TH-YOLOv5 for object detection and classification in this system, and Deepsort was used for monitoring the human. We calculated the physical distance between two or more peoples using pairwise

L2 normalization to create 3D feature space. To monitor the physical distance, we generate a bounding box with text (safe and unsafe) indicating and monitoring who follows/breaks the actual distance. The number of violations is verified by calculating the number of groups created and the violation index term as the ratio of individuals to groups. The suggested approach for detecting people and estimating physical distances is tested using the Oxford Town Centre data set, including over 7,500 people detection and distance estimates. We also used two pedestrian benchmarks data sets to evaluate the proposed system, including HumanCrowd and Citypersons. The system performed well under various conditions, including occlusion, illumination changes, and partial vision. It demonstrated a significant improvement in the mAP score of 0.895 and the speed of 29 FPS, which comparatively outperforms the well-known recent state of art object detectors. In the future, this technique will be applied to mobile cameras, such as those placed on autonomous drones. Drones are thus easier to control and more effective at tracking fast-moving objects in all directions.

Acknowledgements This work was supported in part by the Scientific and Technological Research Council of Turkey (TUBITAK) through the 2232 Outstanding International Researchers Program under Project No. 118C301.

Declarations

Conflict of interest We (authors) certify that there is no actual or potential conflict of interest related to this article.

Compliance with Ethical Standards This article does not contain any studies with human participants and/or animals performed by any of the authors

References

- Worldometer: COVID-19 CORONAVIRUS PANDEMIC. https://www.worldometers.info/coronavirus/?utm_campaign=homeAdvegas1? [Online; accessed 18-June-2021] (2021)
- guardian T. Delta variant of Covid spreading rapidly and detected in 74 countries. 2021. <https://www.theguardian.com/world/2021/jun/14/>. [Online; accessed 25-June-2021]
- AI L. Landing AI Creates an AI Tool to Help Customers Monitor Social Distancing in the Workplace. 2021. <https://landing.ai/>. [Online; accessed 07-June-2021]
- Ksentini A, Brik B. An edge-based social distancing detection service to mitigate covid-19 propagation. *IEEE Internet of Things Magazine*. 2020;3(3):35–9.
- Pouw CA, Toschi F, van Schadowijk F, Corbetta A. Monitoring physical distancing for crowd management: Real-time trajectory and group analysis. *PLoS ONE*. 2020;15(10):0240963.
- Rezaei M, Azarmi M. Deepsocial: Social distancing monitoring and infection risk assessment in covid-19 pandemic. *Appl Sci*. 2020;10(21):7514.
- Saponara S, Elhanashi A, Gagliardi A. Implementing a real-time, ai-based, people detection and social distancing measuring system for covid-19. *Journal of Real-Time Image Processing*, 1–11. 2021.
- Razavi M, Alikhani H, Janfaza V, Sadeghi B, Alikhani E. An automatic system to monitor the physical distance and face mask wearing of construction workers in covid-19 pandemic. *SN computer science*. 2022;3(1):1–8.
- Bentafat E, Rathore MM, Bakiras S. A practical system for privacy-preserving video surveillance. In: *International Conference on Applied Cryptography and Network Security*, 2020:21–39. Springer
- Das S, Nag A, Adhikary D, Ram RJ, BR A, Ojha SK, Hegde GM. Computer vision-based social distancing surveillance solution with optional automated camera calibration for large scale deployment. 2021. arXiv preprint [arXiv:2104.10891](https://arxiv.org/abs/2104.10891)
- Nguyen CT, Saputra YM, Van Huynh N, Nguyen N-T, Khoa TV, Tuan BM, Nguyen DN, Hoang DT, Vu TX, Dutkiewicz E, et al. A comprehensive survey of enabling and emerging technologies for social distancing-part ii: Emerging technologies and open issues. *IEEE Access*. 2020;8:154209–36.
- Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric. In: *2017 IEEE International Conference on Image Processing (ICIP)*, 2017:3645–3649. IEEE
- Suresh K, Bhuvan S, Palangappa M. Social distance identification using optimized faster region-based convolutional neural network. In: *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, 2021:753–760. IEEE
- Ahmed F, Zviedrite N, Uzicanin A. Effectiveness of workplace social distancing measures in reducing influenza transmission: a systematic review. *BMC Public Health*. 2018;18(1):1–13.
- Bouhlef F, Mliki H, Hammami M. Crowd behavior analysis based on convolutional neural network: Social distancing control covid-19. In: *VISIGRAPP (5: VISAPP)*, 2021:273–280
- Razavi M, Alikhani H, Janfaza V, Sadeghi B, Alikhani E. An automatic system to monitor the physical distance and face mask wearing of construction workers in covid-19 pandemic. 2021. arXiv preprint [arXiv:2101.01373](https://arxiv.org/abs/2101.01373)
- Rahim A, Maqbool A, Rana T. Monitoring social distancing under various low light conditions with deep learning and a single motionless time of flight camera. *PLoS ONE*. 2021;16(2):0247440.
- Khandelwal P, Khandelwal A, Agarwal S, Thomas D, Xavier N, Raghuraman A. Using computer vision to enhance safety of workforce in manufacturing in a post covid world. 2020. arXiv preprint [arXiv:2005.05287](https://arxiv.org/abs/2005.05287)
- supply L. Landing AI Creates an AI Tool to Help Customers Monitor Social Distancing in the Workplace. 2021. <https://levelfivesupplies.com/social-distance-monitoring/>. [Online; accessed 02-June-2021]
- Punn NS, Sonbhadra SK, Agarwal S, Rai G. Monitoring covid-19 social distancing with person detection and tracking via fine-tuned yolo v3 and deepsort techniques. 2020. arXiv preprint [arXiv:2005.01385](https://arxiv.org/abs/2005.01385)
- Jocher G, Stoken A, Borovec J, Chaurasia A, Xie T, Liu C, Abhiram V, Laughing T, et al. ultralytics/yolov5: v5. 0-yolov5-p6 1280 models. AWS, Supervise. ly and YouTube integrations **10**. 2021
- Junayed MS, Islam MB, Imani H, Aydin T. Pds-net: A novel point and depth-wise separable convolution for real-time object detection. *International Journal of Multimedia Information Retrieval*. 2022;11(2):171–88.
- Liu S, Qi L, Qin H, Shi J, Jia J. Path aggregation network for instance segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018:8759–8768
- Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016:779–788
- Bochkovskiy A, Wang C-Y, Liao H-YM. Yolov4: Optimal speed and accuracy of object detection. 2020. arXiv preprint [arXiv:2004.10934](https://arxiv.org/abs/2004.10934)
- Shao S, Zhao Z, Li B, Xiao T, Yu G, Zhang X, Sun J. Crowdhuman: A benchmark for detecting human in a crowd. 2018. arXiv preprint [arXiv:1805.00123](https://arxiv.org/abs/1805.00123)
- Zhang S, Benenson R, Schiele B. Citypersons: A diverse dataset for pedestrian detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017:3213–3221
- Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al. An image is worth 16x16 words: Transformers for image recognition at scale. 2020. arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929)
- Woo S, Park J, Lee J-Y, Kweon IS. Cbam: Convolutional block attention module. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018:3–19
- Wojke N, Bewley A. Deep cosine metric learning for person re-identification. In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2018:748–756. IEEE
- Gloude-mans D, Gloude-mans N, Abkowitz M, Barbour W, Work DB. Quantifying social distancing compliance and the effects of behavioral interventions using computer vision. In: *Proceedings of the Workshop on Data-Driven and Intelligent Cyber-Physical Systems*. 2021:1–5
- Sengupta K, Srivastava PR. Hrnet: Ai-on-edge for mask detection and social distancing calculation. *SN Computer Science*. 2022;3(2):1–15.
- Rezaei M, Klette R. *Computer vision for driver assistance*. Cham: Springer International Publishing **45**. 2017.
- Pias: Object detection and distance measurement. 2021. <https://github.com/paul-pias/Object-Detection-and-Distance-Measurement/>. [Online; accessed 11-March-2021]
- Khan MA, Paul P, Rashid M, Hossain M, Ahad MAR. An ai-based visual aid with integrated reading assistant for the completely blind. *IEEE Transactions on Human-Machine Systems*. 2020;50(6):507–17.
- Chen X, Fang H, Lin T-Y, Vedantam R, Gupta S, Dollár P, Zitnick CL. Microsoft coco captions: Data collection and evaluation server. 2015. arXiv preprint [arXiv:1504.00325](https://arxiv.org/abs/1504.00325)
- Harvey A, LaPlace J. Megapixels: origins, ethics, and privacy implications of publicly available face recognition image datasets. *Megapixels*. 2019.

38. Ren S, He K, Girshick R, Sun J. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Mach Intell.* 2016;39(6):1137–49.
39. Redmon J, Farhadi A. Yolov3: An incremental improvement. 2018. arXiv preprint [arXiv:1804.02767](https://arxiv.org/abs/1804.02767)
40. Tan M, Pang R, Le QV. Efficientdet: Scalable and efficient object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2020:10781–10790.
41. Yang D, Yurtsever E, Renganathan V, Redmill K, Özgüner Ü. A vision-based social distance and critical density detection system for covid-19. 2020.
42. Yadav S, Gulia P, Gill NS, Chatterjee JM. A real-time crowd monitoring and management system for social distance classification

and healthcare using deep learning. *Journal of Healthcare Engineering* **2022.** 2022.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.