

DeepSOCIAL: Social Distancing Monitoring and Infection Risk Assessment in COVID-19 Pandemic

Mahdi Rezaei ^{1,*}, Mohsen Azarmi ^{2,*}

¹ Institute for Transport Studies, The University of Leeds, Leeds, LS2 9JT, UK

² Department of Computer Engineering, Qazvin Azad University, Qazvin, IR

¹ m.rezaei@leeds.ac.uk ² m.azarmi@qiau.ac.ir

ABSTRACT

Social distancing is a recommended solution by the World Health Organisation (WHO) to minimise the spread of COVID-19 in public places. The majority of governments and national health authorities have set the 2-meter physical distancing as a mandatory safety measure in shopping centres, schools and other covered areas. In this research, we develop a generic Deep Neural Network-Based model for automated people detection, tracking, and inter-people distances estimation in the crowd, using common CCTV security cameras. The proposed model includes a YOLOv4-based framework and inverse perspective mapping for accurate people detection and social distancing monitoring in challenging conditions, including people occlusion, partial visibility, and lighting variations. We also provide an online risk assessment scheme by statistical analysis of the Spatio-temporal data from the moving trajectories and the rate of social distancing violations. We identify high-risk zones with the highest possibility of virus spread and infection. This may help authorities to redesign the layout of a public place or to take precaution actions to mitigate high-risk zones. The efficiency of the proposed methodology is evaluated on the Oxford Town Centre dataset, with superior performance in terms of accuracy and speed compared to three state-of-the-art methods.

Keywords – Social Distancing; COVID-19; People Detection and Tracking; Inverse Perspective Mapping; Deep Learning; Convolutional Neural Networks.

1 Introduction

THE novel generation of the coronavirus disease (COVID-19) were reported in late December 2019 in Wuhan, China. After only a few months, the virus was hit by the global outbreak in 2020. On May 2020 The World Health Organisation (WHO) announced the situation as pandemic^{1,2}. The statistics by WHO on 26 August 2020 confirms 23.8 million infected people in 200 countries. The mortality rate of the infectious virus also shows a scary number of 815,000 people.

With the growing trend of patients, there is still no effective cure or available treatment for the virus. While scientists, healthcare organisations, and researchers are continuously working to produce appropriate medications or vaccines for the deadly virus, no definite success has been reported at the time of this research, and there is no certain treatments or recommendation to prevent or cure this new disease. Therefore, precautions are taken by the whole world to limit the spread of infection. These harsh conditions have forced the global communities to look for alternative ways to reduce the spread of the virus.

Social distancing, as shown in Figure 1(a), refers to precaution actions to prevent the proliferation of the disease, by minimising the proximity of human physical contacts in covered or crowded public places (e.g. schools, workplaces,



(a) Social distancing monitoring



(b) Accumulated infection risk (red-spots) due to multiple breaches of the social-distancing

Figure 1. People Detection, Tracking, and Risk assessment in Oxford Town Centre, using a public CCTV camera.

gyms, lecture theatres, etc.) to stop the widespread accumulation of the infection risk (Figure 1(b)).

According to the defined requirements by the WHO, the minimum distance between individuals must be at least 6 feet (1.8 meters) in order to observe an adequate social distancing

 Corresponding Author: m.rezaei@leeds.ac.uk (M. Rezaei)

* Both authors contributed equally

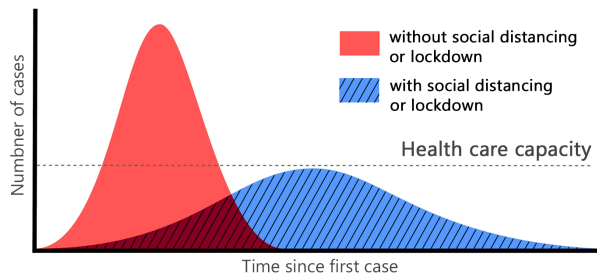


Figure 2. Gaussian distribution of infection transmission rate for a given population, with and without social distancing obligation.

among the people³.

Recent researches have confirmed that people with mild or no symptoms may also be carriers of the novel Coronavirus infection⁴. Therefore, it is important all individuals maintain controlled behaviours and observe social distancing.

Many research works such as^{5,6,7} have proved social-distancing as an effective non-pharmacological approach and an important inhibitor for limiting the transmission of contagious diseases such as H1N1, SARS, and COVID-19.

Figure 2 demonstrates the effect of following appropriate social distancing guideline to reduce the rate of infection transmission among individuals^{8,9}. A wider Gaussian curve with a shorter spike within the range of the health system service capacity makes it easier for patients to fight the virus by receiving continuous and timely support from the health care organisations. Any unexpected sharp spike and rapid infection rate (such as the red curve in Figure 2), will lead to the service failure, and consequently, exponential growth in the number of fatalities.

For several months, the World Health Organisation believed that COVID-19 is only transmittable via droplets emitted when people sneeze or cough and the virus does not linger in the air. However, on 8 July 2020, the WHO announced:

“There is emerging evidence that COVID-19 is an airborne disease that can be spread by tiny particles suspended in the air after people talk or breathe, especially in crowded, closed environments or poorly ventilated settings.”

Therefore, social distancing now claims to be even more important than thought before, and one of the best ways to stop the spread of the disease in addition to wearing face masks. Almost all countries are now considering it as a mandatory practice.

During the COVID-19 pandemic, governments have tried to implement a variety of social distancing practices, such as restricting travels, controlling borders, closing pubs and bars, and alerting the society to maintain a distance of 1.6 to 2 meters from each other¹⁰. However, monitoring the amount of infection spread and efficiency of the constraints is not an easy task. People require to go out for essential needs such as

food, health care and other necessary tasks and jobs.

In such situations, Artificial Intelligence can play an important role in facilitating social distancing monitoring. Computer vision, as a subset of Artificial Intelligence, has been very successful in solving various complex health care problems and has shown its potential in Chest CT-Scan or X-Ray based COVID-19 recognition^{11,12}. Besides, deep neural networks enable us to extract complex features from the data so that we can provide a more accurate understanding of the images by analysing and classifying these features. Examples include diagnosis, clinical management and treatment, as well as the prevention and control of COVID-19^{13,14}.

The main contribution of this research can be highlighted as follows:

- This study aims to support the reduction of the corona virus spread and its economic costs by providing an AI-based solution to automatically monitor and detect violations of social distancing among individuals.
- We develop one of the most (if not the most) accurate deep neural network (DNN) models for people detection, tracking, and distance estimation called *DeepSOCIAL*.
- We perform a live and dynamic risk assessment, by statistical analysis of spatio-temporal data from the people movements at the scene.
- The developed model is a generic human detection and tracker, not limited to social-distancing monitoring, and can be applied for various real-world applications such as pedestrian detection in autonomous vehicles, human action recognition, anomaly detection, and security systems.

More details and further information will be provided in the following sections. In Section 2 we discuss the related work, existing challenges, and research gaps in the field. The proposed mythology including the model architecture and our object detection techniques, tracking, and red-zone prediction algorithm will be proposed in Section 3. In Section 4 experimental results and performance of the system is investigated against the state-of-the-art, followed by discussions and concluding remarks in Section 5.

2 Related Works

Many researchers have worked in the medical and pharmaceutical fields aiming at treatment of COVID-19 infectious disease; however, no definite solution has yet been found. On the other hand, controlling the spread of such an unknown respiratory infectious disease is another issue.

Variety of studies with different implementation strategies^{5,6,15} have proven that controlling the prevalence is a contributing factor, and social distancing is an effective way to reduce the transmission and prevent the spread of the virus in society.

Several researchers such as¹⁶ and¹⁵ use SIR (Susceptible, Infectious, or Recovered) model. SIR is an epidemiological modelling system to compute the theoretical number of people infected with a contagious disease in a given population, over time. One of the oldest yet common SIR models is Kermack and McKendrick models introduced in 1927¹⁷. Eksin et al.¹⁶, have recently introduced a modified model of SIR by including a social distancing parameter, which can be used to determine the number of infected and recovered individuals.

Effectiveness of social distancing practices can be evaluated based on several standard approaches. One of the main criteria is based on the reproduction ratio, R_o , which indicates the average number of people who may be infected from an infectious person during the entire period of the infection¹⁸. Any $R_o > 1$ indicates an increasing rate of infection within the society and $R_o < 1$ indicates that every case will infect less than 1 person, hence, the disease rate is considered to be declining in the target population.

Since the R_o value indicates the disease outbreak, it is one of the most important indicators for selecting social distancing criteria. In the current COVID-19 pandemic, the World Health Organisation estimated the R_o rate would be in the range of 2-2.5¹⁹, which is significantly higher than other similar diseases such as seasonal flu with $R_o = 1.4$. In²⁰, a clear conclusion is drawn about the importance of applying social distancing for the cases with a high amount of R_o .

In another research based on the game theory on the classic SIR model, an assessment of the benefits and economic costs of social distancing has been examined²¹. The results also show that in the case of $R_o < 1$, social distancing would cause unnecessary costs, while $R_o \approx 2$ implies that social distancing measures have the highest economic benefits. In a similar research, Kylie et al.²² investigated the relationship between the stringency of social distancing and the region's economic status. This study suggests although preventing the widespread outbreak of the virus is necessary, a moderate level of social activities could be allowed.

Prem et al.²³ use location-specific contact patterns to investigate the effect of social distancing measures on the prevalence of COVID-19 pandemic in order to remove the persistent path of disease outbreak using susceptible-exposed-infected-removed models (SEIR).

Since the onset of coronavirus pandemic, many countries have used technology-based solutions, to inhibit the spread of the disease^{24, 25, 26}. For example, some of developed countries, such as South Korea and India, use GPS data to monitor the movements of infected or suspected individuals to find any possible exposure among the healthy people. The India government uses the *Aarogya Setu* program to find the presence of COVID-19 patients in the adjacent region, with the help of GPS and Bluetooth. This may also help other people to maintain a safe distance from the infected person²⁷. Some law enforcement agencies use drones and surveillance cameras to detect large-scale rallies and have carried out regulatory measures to disperse the population^{28, 29}. Some other researchers

such as Xin et al.³⁰ perform human detection using wireless signals by identifying phase differences and change detection in amplitude wave-forms. However this requires multiple receiving antennas and can not be easily integrated in all public places.

The utilisation of Artificial Intelligence, Computer Vision, and Machine Learning, enables us to discover the correlation of high-level features. For example, it enables us to understand and predict pedestrian behaviours in traffic scenes, sports actions and activities, medical imaging, anomaly detection, etc. by analysing Spatio-temporal visual information and statistical data analysis of the images sequences^{31, 14}.

In health-related fields, it would be also feasible to predict the sickness trend of specific areas³², to predict the density of people in public places³³, or determine the distance of individuals from the popular swarms³⁴, using a combination of visual and Geo-location cellular information. However, such research works suffer from challenges such as skilled labour or the cost of designing and implementing the infrastructures.

On the other hand, recent advances in AI, Computer Vision, Deep Learning, and Pattern Recognition, enables the computers to understand and interpret the visual data from digital images or videos. It also allows computers to identify and classify different types of objects^{35, 36, 37}. Such capabilities can play an important role in empowering, encouraging, and performing social distancing surveillance and measurements as well. For example, computer vision could turn CCTV cameras in the current infrastructure capacity into "smart" cameras that not only monitor people but can also determine whether people follow the social distancing guidelines or not. Such systems require a very precise human detection algorithms.

People detection in image sequences is one of the most important sub-branches in the field of object detection and computer vision. Although many research works have been done in the field of human detection³⁸ and human action recognition³⁹, majority of them are either limited to indoor applications or suffer from accuracy issues under outdoor challenging lighting conditions. A range of other research rely on manual tuning methodologies to identify people activities, however, limited functionality has always been an issue⁴⁰.

Convolutional Neural Networks (CNN) have played a very important role in feature extraction and complex object classification. With the development of faster CPUs, GPUs, and extended memory capacities, CNNs allow the researchers to make accurate and fast detectors compared to conventional models. However, the long time training, detection speed and achieving better accuracy, are still remaining challenges to be solved.

Narinder et al.⁴¹ used a deep neural network (DNN) based detector, along with DeepSort⁴² algorithm as an object tracker for people detection to assess the distance violation index. However, no statistical analysis of the outcome of their results is provided. Furthermore, no discussion about the validity of the distance measurements is provided.

In another study by Prateek et al.⁴³, the authors have ad-

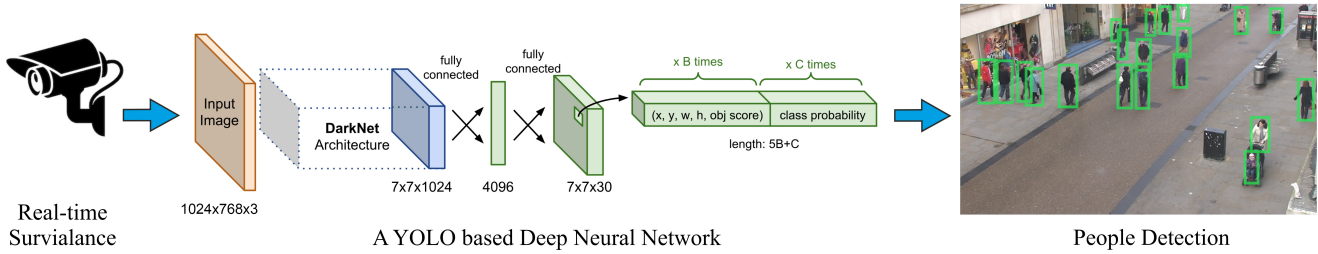


Figure 3. Stage 1- The overall structure of the people detection module.

addressed the people distancing in a given manufactory. They have used MobileNet V2 network⁴⁴ as a lightweight detector to reduce computational costs, which in turn provides less accuracy comparing to some other common models.

Similar to the other research, no statistical analysis is performed on the results of the distance measurement in⁴⁵. The authors have made a comparison between two common types of DNN models (YOLO and Faster R-CNN); however, the system has been only tested on a basic dataset.

Since the topic is very recent, there has not been much research regarding the accuracy of detections, no experiment on challenging datasets has been performed, no standard comparison has been conducted on common datasets, and no analytical studies or post-processing have been considered after the detection-phase.

Considering the above-mentioned research gaps, we propose a new model which not only performs more accurate and faster than the state-of-the-art but also will be trained and tested using large and comprehensive datasets, in challenging environments and lighting conditions. This will ensure the model is capable of performing in real-world scenarios, particularly in covered shopping centres where the lighting conditions are not as ideal as the outdoor lighting. Furthermore, we offer post-detection and post-processing analytical solutions to mitigate the spread of the virus.

3 Methodology

We propose a 3-stage model including people detection, tracking, inter-distance estimation as a total solution for social distancing monitoring and zone-based infection risk analysis. The system can be integrated and applied on all types of CCTV surveillance cameras with any resolution from VGA to Full-HD, with real-time performance.

3.1 People Detection

Figure 3 shows the overall structure of the Stage 1. The objective is to develop a model to detect humans (people) with various types of challenges such as variations in clothes, postures, at far and close distances, with/without occlusion, and under different lighting conditions.

To gain this we inspire from the strength of cutting-edge research; however, we develop our own unique human classifier and train our model based on a set of comprehensive and multifaceted datasets. Before diving into further technical details,

we overview the most advanced object detection techniques and then introduce our human detection model.

Modern DNN-based detectors consist of two sections: A *backbone* for extracting features and a *head* for predicting classes and location of objects. The feature extractor tends to encode model inputs by representing specific features that help it to learn and discover the relevant patterns related to the query object(s). Examples of feature extraction architectures can be seen in VGG16⁴⁶, ResNet-50⁴⁷, CSPResNeXt-50⁴⁸, CSPDarknet53⁴⁸, and EfficientNet-B0/B7⁴⁹. The head of a DNN is responsible for classifying the objects (e.g. people, bicycles, chairs, etc.) as well as calculating the size of the objects and the coordinates of the correspondent bounding boxes.

There are usually two types of head sections: one-stage and two-stage. The two-stage detectors use the region proposal before applying the classification. First, the detector extracts a set of object proposals (candidate bounding boxes) by a selective search. Then it resizes them to a fixed size before feeding them to the CNN model. This is similar to R-CNN based detectors⁵⁰⁻⁵². There are usually two types of head sections: one-stage and two-stage. The two-stage detectors use the region proposal before applying the classification. First, the detector extracts a set of object proposals (candidate bounding boxes) by a selective search. Then it resizes them to a fixed size before feeding them to the CNN model. This is similar to R-CNN based detectors⁵⁰⁻⁵². In spite of the accuracy of two-stage detectors, such methods are not suitable for the systems with restricted computational resources⁵³.

On the other hand, the one-stage detectors perform a single detection process such as the work done by Liu et al. (known as SSD), or other works done by Redmon et al.^{54, 55, 56}, and Bochkovski et al⁵⁷, known as “You Only Look Once” or YOLO detectors. Such detectors use regression analysis to calculate the dimensions of bounding boxes and interpret their class probabilities. It maps the image pixels to the enclosed grids and checks the probability of the existence of an object in each cell of the grids. This approach offers excellent improvements in terms of speed and efficiency.

Figure 4, shows the outcome of our investigations and reviews on some of the most successful object detection models such as RCNN⁵⁰, fast RCNN⁵¹, faster RCNN⁵², SSD⁵⁸, YOLOv1-v4⁵⁴⁻⁵⁷ tested on the PASCAL-VOC⁵⁹ and MS-COCO⁶⁰ data sets under similar conditions.

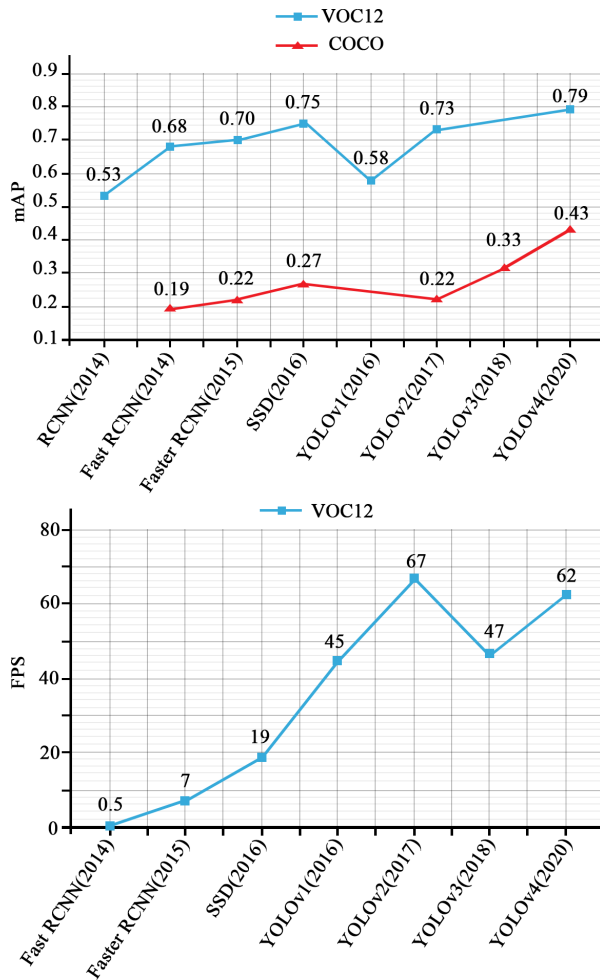


Figure 4. Speed and Precision overview of eight most popular object detection models on MS-COCO and PASCAL-VOC datasets.

Otherwise, the performance of the systems may vary depending on various factors such as backbone architecture, input image size, resolution, model depth, software, and hardware platform.

As illustrated in Figure 4, YOLOv4 offers the best trade-off for the speed and the accuracy; however, since YOLOv4 is an aggregation of various techniques, we undertook an in-depth study of each sub-techniques to achieve the best results for our people detection model and to outperform the state-of-the-art.

We considered two approaches to improve the accuracy of the backbone: A basic way to improve the accuracy of CNN-based detectors is to expand the receptive field and enhance the complexity of the model using more layers; however, using this technique makes it harder to train the model. We suggest using a skip-connections technique for the ease of training, instead.

Various models use a similar policy to make connections between layers, such as Cross-Stage-Partial-connections (CSP)⁴⁸ or Dense Blocks (consisting of Batch Normalisation, ReLU, Convolution, etc.) in DenseNet⁶¹. Such models

Table 1. Comparisons of three backbone models in terms of number of parameters and speed (*fps*) using an RTX 2070 GPU.

Backbone model	Input Resolution	Number of Parameters	Speed (<i>fps</i>)
CSPResNeXt50	512×512	20.6 M	62
CSPDarknet53	512×512	27.6 M	66
EfficientNet-B3	512×512	12.0 M	26

have also been used in the design of some recent backbone architectures, such as CSPResNeXt50, CSPDarknet53⁴⁸ and EfficientNet-B3, which are our supported architectures options for YOLOv4.

Table 1 summarises the brief report of our investigations for the above-mentioned backbone architectures in terms of number of parameters and the processing speed (in *fps*) for the same input size of 512×512 .

Based on theoretical justifications in⁵⁷ and several experiments by us, we concluded that CSPDarknet53 is the most optimal backbone model for our application, in spite of higher complexity (due to more number of parameters). Here, the higher number of parameters leads to the increased capability of the model in detecting multiple objects while at the same time we can maintain real-time performance.

Recently, some of the modern proposed models have placed some extra layers between the backbone and the head, called the *neck*, which is considered for feature collection from different stages of the backbone network.

The neck section consists of several top-down and bottom-up paths to collect and combine parameters of the network in different layers, in order to provide a more accurate image features for the head section. Some of the techniques used in the neck section are the Feature Pyramid Network (FPN)⁶², Path Aggregation Network (PAN)⁶³ and Spatial Attention Module (SAM)⁶⁴.

In YOLO-v4, the authors have dealt with two categories of training options for different parts of the network: “*Bag of Freebies*”, which includes a set of methods to alter the model’s training strategy with the aim of increasing the generalisation; and “*Bag of Specials*” which includes a set of modules that can significantly improve the object detection accuracy in exchange for a small increase in training costs.

Many CNN-based models use Fully-Connected layers for classification part, and consequently, they can only accept fixed dimensions of images as input. This can lead to two types of issues: firstly, we cannot deal with low resolution images and secondly, the detection of the small objects would be difficult. These are in contradiction with our objectives where we aim to have our model applicable in any surveillance cameras with any input image sizes and resolutions.

In order to deal with the first issue we can refer to existing methodologies such as Fully Convolution Networks (FCN). These has no FC-layers and therefore can deal with images

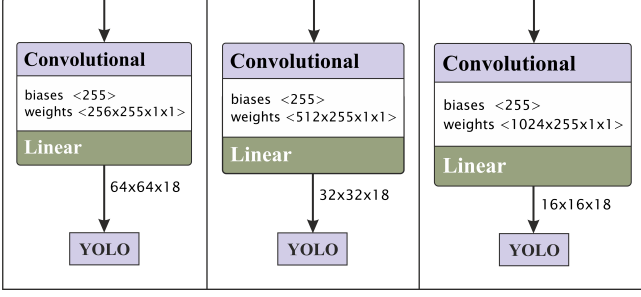


Figure 5. The YOLO-based heads applied at different scales

with different sizes. However, to cope with the second issue (i.e. dealing with small objects), we performed a feature pyramid technique to enhance the receptive field and extract different scales of image from the backbone, and finally, performing a multi-scale detection in the head section.

Figure 5 shows the heads that we applied at different scales of the network for object detection at different sizes. YOLOv3 uses FPN⁶² to extract features of different scales from the backbone, however in YOLOv4 based approach we use a modified version of the Spatial Pyramid Pooling layer (SPP)⁶⁵, instead of FPN, to address the issue of various spatial dimensions as well as dealing with multi-scale detection in head section.

In SPP module we see the input image as a “Bag of Words” feature-map which is divided into $m \times m$ bins, where m can be 1, 2, or 4. A max-pooling is then applied to each bin for each channel to get best features. In section 4 we will examine the efficiency of this approach in improving the accuracy of our model.

In DNNs, the bottom layers (i.e the first few layers) extract localised pattern and texture information to gradual build up of the semantic information which is required in the top layers. However, during the feature extraction process, some parts of the local information that may be required for fine-tuning of the model, may be lost. In PANet⁶³ approach, the information flow of the bottom layers will be added to top layers to strengthen of localised information; therefore, better fine-tuning and prediction can be expected. In a recent research by Bochkovskiy et al.⁵⁷, it is shown that the concatenation operator performs better than addition operator to maintain the localised information and transferring them to the top layers.

Experimenting various rational configurations for the neck module of our model, we use spatial pyramid pooling (SPP) and PAN as well as the Spatial Attention Module (SAM)⁶⁴ which together made one of the most effective, consistent and robust components to focus the model on optimising the parameters.

In the head of the model, we use the same configuration as YOLOv3. Similar to many other anchor-based models, YOLO uses predefined boxes to detect multiple objects. Then the object detection model will be trained to predict each generated anchor boxes that belong to a particular class. After that, an offset will be used to adjust the dimensions of the

anchor box in order to better match with the ground-truth data, based on the classification and regression loss.

Assuming the grid cell reference point (c_x, c_y) at the top left corner of the object image and the bounding box prior with the width and height (p_w, p_h) , the network predicts a bonding box at the centre (\hat{x}, \hat{y}) and the size of (\hat{w}, \hat{h}) with the corresponding offset and scales of (b_x, b_y, b_w, b_h) as follows:

$$\begin{aligned} \hat{x} &= \sigma(b_x) + c_x \\ \hat{y} &= \sigma(b_y) + c_y \\ \hat{w} &= p_w e^{b_w} \\ \hat{h} &= p_h e^{b_h} \end{aligned} \quad (1)$$

where σ is the Sigmoid confidence score function within the range of 0 and 1.

We represent the class of “human” with a 4-tuple (x, y, w, h) , where (x, y) is the centre of the bounding box, and w, h are width and height, respectively.

We use three anchor boxes to find a maximum of three people in each grid cell. Therefore, the total number of channels is 18: (1 class + 1 object + 4 coordinates) \times 3 anchors.

Since we have multiple anchor boxes for each spatial location, a single object may be associated with more than one anchor boxes. This problem can be resolved by using non-maximal suppression (NMS) technique and by computing intersection over union (IoU) to limit the anchor boxes association.

As part of the weight adjustment and loss minimisation operation, we use Complete IoU (CIoU) (in Eq.3) instead of the basic IoU (Eq. 2). The CIoU not only compares the location and distance of the candidate bounding boxes to the ground truth bounding box but also it compares the aspect ratio of the size of the generated bounding boxes with the size of the ground truth bounding box.

$$IoU = \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|} \quad (2)$$

$B^{gt} = (x^{gt}, y^{gt}, w^{gt}, h^{gt})$ is the ground-truth box, and $B = (x, y, w, h)$ is the predicted box. We use CIoU not only as a detection metric but also as a loss function:

$$\mathcal{L}_{CIoU} = 1 - IoU + \frac{|\rho^2(B, B^{gt})|}{c^2} + \alpha v \quad (3)$$

where ρ is Euclidean distance between grand-truth B^{gt} , and predicted B bounding box. The diagonal length of the smallest bounding box enclosing both boxes B and B^{gt} is represented by c ; and α is a positive trade-off parameter:

$$\alpha = \frac{v}{(1 - IoU) + v} \quad (4)$$

and v measures the consistency of the aspect ratio, as follows:

$$v = \frac{4}{\pi^2} \left(\arctan \frac{w^{gt}}{h^{gt}} - \arctan \frac{w}{h} \right)^2 \quad (5)$$

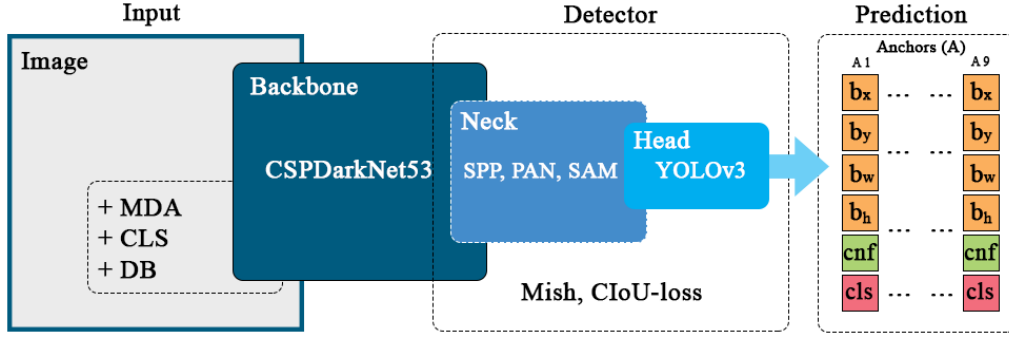


Figure 6. The network structure of the proposed 3-level human detection module.

Even in case of zero-percent overlapping, the loss functions still gives us an indication on how to adjust the weights to firstly converge the aspect size towards 1, and secondly, how to reduce the error distance of candidate bounding boxes to the centre of the ground truth bounding box. A similar approach called Distance-IoU is used in⁶⁶ for another application.

We also investigated the performance of our model against three activation functions.

ReLU:

$$f(x) = \max(0, x) \quad (6)$$

$$f'(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

Swish:

$$f(x) = x \cdot \frac{1}{(1 + e^{-x})} \quad (8)$$

$$f'(x) = \frac{1}{(1 + e^{-x})} \cdot \left(1 - \frac{1}{(1 + e^{-x})}\right) \quad (9)$$

and **Mish:**

$$\begin{aligned} f(x) &= x \cdot \tanh(\text{softplus}(x)) \\ &= x \cdot \tanh(\ln(1 + e^x)) \end{aligned} \quad (10)$$

$$f'(x) = \frac{e^x \omega}{\delta^2} \quad (11)$$

as a self regularised Non-Monotonic activation function, where

$$\omega = 4(x + 1) + 4e^{2x} + e^{3x} + (4x + 6) \quad (12)$$

and

$$\delta = 2e^x + e^{2x} + 2. \quad (13)$$

Our preliminary evaluations, confirmed the same results provided by Misra⁶⁷ for our human detection application. The Mish activation function converged towards the minimum loss, faster than Swish and ReLU, with higher accuracy. The result was consistent especially for diversity of parameter initialisers, regularisation methods, and lower learning rate values.

Among various techniques of Bag of Freebies, we used the Mosaic data augmentations⁵⁷ which integrates four images into one, to increase the size of the input data without requiring to increase the batch size.

On the other hand, in batch normalisation, the batch size reduction causes noisy estimation of mean and variance. To address this issue, we considered the normalised values of the previous k iterations instead of a single mini-batch. This is similar to Cross-Iteration Batch Normalisation (CBM)⁶⁸.

As mentioned in⁶⁹, using Dropblock regularisation can effectively prevent overfitting. In this regard, Class label Smoothing⁷⁰ also helps to prevent over-fitting by reducing the model's confidence in the training phase.

Figure 6 shows the 3-level structure of the human detection module and the sequence of the network components. In the input part, the Mosaic data augmentation (MDA), Class label smoothing (CLS) and DropBlock regularisation are applied to the input image. In the detector part, the Mish activation function has been used, and CIoU metric is considered as the loss function. In the prediction part, for each cell at each level, the anchor boxes contain the information required to locate the bounding box, confidence ratio of the object, and the corresponding class of the object. In total, we have 9 anchor boxes.

We also used Stochastic Gradient Descent (SGD) with *warm restarts*⁷¹ to change the learning rate during the training process. This helps to jump out of local minima in the solution space and save the training time. The method initially considers a large value of the learning rate, then slows down the learning speed halfway, and eventually reduces the learning rate for each batch, with a tiny downward slope. We decreased the learning rate using a cosine annealing function for each batch as follows:

$$\eta_t^i = \eta_{min}^i + \frac{1}{2}(\eta_{max}^i - \eta_{min}^i)(1 + \cos(\frac{T_{cur}}{T_i} \pi)) \quad (14)$$



Figure 7. People Detection, ID assignment, Tracking and moving trajectory representation.

where η_t is the current learning rate in the i^{th} run, η_{min} and η_{max} are the minimum and maximum target learning rates. T_{cur} is the number of epochs executed since the last restart, and T_i is the number of epochs performed since the restart of the SGD.

For the training phase of the model, we considered the *person* category of the Microsoft COCO Dataset⁶⁰ as well as the Open Images Datasets V6+⁷² which includes 16 Million ground truth bounding boxes in 600 categories. The dataset is a collection of 19,957 classes and the major part of the dataset is suitable for human detection and identification. The dataset is annotated using the bounding-box labels on each image along with the corresponding coordinates of each label.

We also considered the category of human body parts such as the legs as we believe this allows the detector to learn a more general concept of a human being, particularly in occluded situations or in case of partial visibility e.g. at the borders of the input image where the full-body of the individuals can not be perceived.

3.2 People Tracking

The next step after the detection phase, is people tracking and ID assignment for each individual.

We use the *Simple Online and Real-time* (SORT) tracking technique⁷³ as a framework for the Kalman filter⁷⁴ along with the *Hungarian* optimisation technique to track the people. Kalman filter predicts the position of the human at time $t + 1$ based on the current measurement at time t and the mathematical modelling of the human movement. This is an effective way to keep localising the human in case of occlusion.

The Hungarian algorithm is a combinatorial optimisation algorithm that helps to assign a unique ID number to identify a given object in a set of image frames, by examining whether a person in the current frame is the same detected person in the previous frames or not.

Figure 7(a) shows a sample of the people detection and ID assignment, Figure 7(b) illustrates the tracking path of each individual, and Figure 7(c) shows the final position and status of each individual after 100 frames of detection, tracking, and ID assignment. We later use such temporal information for analysing the level of social distancing violations and high-risk zones of the scene.

The state of each human in a frame is modelled as:

$$x = [u, v, s, r, u', v', s']^T \quad (15)$$

where (u, v) represent the horizontal and vertical position of the target bounding box (i.e. the centroid); s denoted the scale (area), and r is the aspect ratio of the bounding box sides.

When an identified human associates with a new observation, the current bounding box will be updated with the newly observed state. This will be calculated based on the velocity and acceleration components, estimated by the Kalman filter framework. If the predicted identities of the query individual significantly differ from the new observation, almost the same state that is predicted by the Kalman filter will be used with almost no correction. Otherwise, the corrections weights will be split proportionally between the Kalman filter prediction and the new observation (measurement).

As mentioned earlier, we use the Hungarian algorithm to solve data association problem, by calculating the IoU (Eq. 2) and the distance (difference) of the actual input values to the predicted values by the Kalman filter.

After the detection and tracking process, for every input frame $I_{w \times h}$ at time t , we define the matrix D_t that includes the location of n detected human in the image carrier grid:

$$D_t = \{P_{(x_n, y_n)}^t \mid x_n \in w, y_n \in h\} \quad (16)$$

3.3 Inter-distance Estimation

Stereo-vision is a popular technique for distance estimation such as in⁷⁵; however, this is not a feasible approach in our research when we aim at integration of an efficient solution, applicable in all public places using only a basic CCTV camera. Therefore we adhere to a monocular solution.

On the other hand, by using a single camera, the projection of a 3-D world scene into a 2-D perspective image plane leads to unrealistic pixel-distances between the objects. This is called perspective effect, in which we can not perceive uniform distribution of distances in the entire image. For example, parallel lines intersect at the horizon and farther people to the camera seem much shorter than the people who are closer to the camera coordinate centre.

In three-dimensional space, the centre or the reference point of each bounding box is associated with three parameters (x, y, z) , while in the image received from the camera, the original 3D space is reduced to two-dimensions of (x, y) , and the depth parameter (z) is not available. In such a lowered-dimensional space, the direct use of the Euclidean distance criterion to measure inter-people distance estimation would be erroneous.

In order to apply a calibrated IPM transition, we first need to have a camera calibration by setting $z = 0$ to eliminate the perspective effect. We also need to know the camera location, its height, angle of view, as well as the optics specifications (i.e. the camera intrinsic parameters)⁷⁴.

By applying the IMP, the 2D pixel points (u, v) will be mapped to the corresponding world coordinate points (X_w, Y_w, Z_w) :

$$\begin{bmatrix} u & v & 1 \end{bmatrix}^T = KRT \begin{bmatrix} X_w & Y_w & Z_w & 1 \end{bmatrix}^T \quad (17)$$

where R is the rotation matrix:

$$R = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta & 0 \\ 0 & \sin\theta & \cos\theta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad (18)$$

T is the translation matrix:

$$T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -\frac{h}{\sin\theta} \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad (19)$$

and K , the intrinsic parameters of the camera are shown by the following matrix:

$$K = \begin{bmatrix} f * ku & s & c_x & 0 \\ 0 & f * kv & c_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \quad (20)$$

where h is the camera height, f is focal length, and ku and kv are the measured calibration coefficient values in horizontal and vertical pixel units, respectively. (c_x, c_y) is the principal point shifts that corrects the optical axis of the image plane.

The camera creates an image with a projection of three-dimensional points in the world coordinate that falls on a retina plane. Using homogeneous coordinates, the relationship between three-dimensional points and the resulting image points of projection can be shown as follows:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \quad (21)$$

where $M \in \mathbb{R}^{3 \times 4}$ is the transformation matrix with m_{ij} elements in Equation 21, that maps the world coordinate points

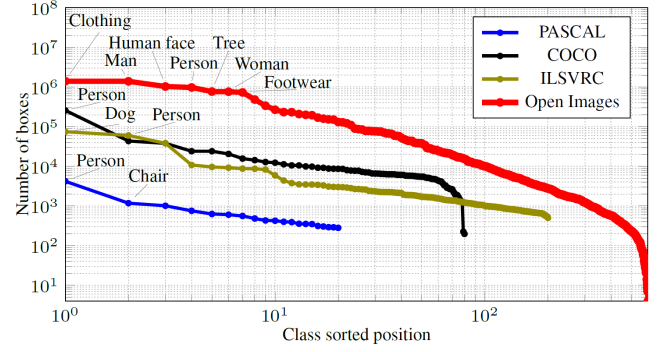


Figure 8. The number of annotated boxes per class in four common datasets. The horizontal axis is represented in logarithmic scale for better readability.

into the image points based on the camera location and the reference frame, provided by the Camera Intrinsic Matrix K (Eq. 20), Rotation Matrix R (Eq. 18) and the Translation Matrix T (Eq. 19).

Considering the camera image plane perpendicular to the Z access in the world coordinate system (i.e. $z = 0$) the dimensions of above equation can be reduced to the following form:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ 1 \end{bmatrix} \quad (22)$$

and finally transferring from the perspective space to inverse perspective space (BEV) can also be expressed in the following scalar form:

$$(u, v) = \left(\frac{m_{11} \times x_w + m_{12} \times y_w + m_{13}}{m_{31} \times x_w + m_{32} \times y_w + m_{33}}, \frac{m_{21} \times x_w + m_{22} \times y_w + m_{23}}{m_{31} \times x_w + m_{32} \times y_w + m_{33}} \right) \quad (23)$$

4 Experimental Results

In order to train the developed model, we considered a transfer learning approach by using pre-trained models on Microsoft COCO dataset⁶⁰ followed by fine-tuning and optimisation of our YOLO-based model.

Four common multi-object annotated datasets including Pascal VOC⁵⁹, COCO⁶⁰, Image Net ILSVRC⁷⁶, and Google Open Images⁷², were investigated in terms of the number of bounding boxes for human or person.

Figure 8 represents the sorted rank of object classes with the number of bounding boxes for each class in each dataset. In Google Open Images dataset (GOI) the class *Person* shows 4th rank, with nearly 10^6 annotated bounding boxes; richer than other three investigated datasets. In addition to the class *Person*, we also adopted four more classes of *Man*, *Woman*,



Figure 9. Examples of annotated images in Open Images dataset.

Boy, and Girl from the GOI dataset for the human training purpose. This made a total number of 3,762,615 samples that we used from training, including 257,253 samples from the COCO dataset and 3,505,362 samples from the GOI dataset.

Figure 9 shows examples of annotated images from the Open Images Dataset. The figure illustrates the diversity of annotated persons including large and small bounding boxes, in far and near distances to the camera image plane, people occlusion, as well as variations in shades and lighting conditions.

4.1 Performance Evaluation

In order to test the performance of the proposed model, we used the Oxford Town Centre (OTC) dataset²⁹ as a previously unseen and challenging dataset with very frequent cases of occlusions, overlaps, and crowded zones. The dataset also contains a good diversity of human specimens in terms of cloths and appearance in a real-world public place.

In order to provide a similar conditions for performance analysis of YOLO based models, we fine-tuned each model on human categories of the Google Open Images (GOI)⁷² data set. This has been done by removing the last layer of each model and placing a new layer (with random values of the uniform probability distribution) corresponding to a binary classification (presence or absence of a human). Furthermore, in order to provide an equal condition for the speed and generalizability, we also tested each of the trained models against the OTC dataset²⁹.

We evaluated and compared our developed models against three common metrics of object detection in computer vision, including Precision Rate, Recall Rate, and FPS against three state-of-the-art human/object detection methods.

All of the benchmarking tests and comparisons were conducted on the same hardware and software: a Windows 10-based platform with an Intel[®] Core™ i5-3570K Processor and an NVIDIA RTX 2080 GPU with CUDA version 10.1.

Figure 10 illustrates the development of loss function in training and validations phases for four versions of our DeepSOCIAL model with different backbone structures. The graphs confirm a fast yet smooth and stable transition for minimising the loss function in DS version after 1090 epochs where we reached to an optimal trade-off point for both the training and validation loss. Table 2 provides the details of each backbone and the outcome of the experimental results against three more state-of-the-art model on the OTC Dataset.

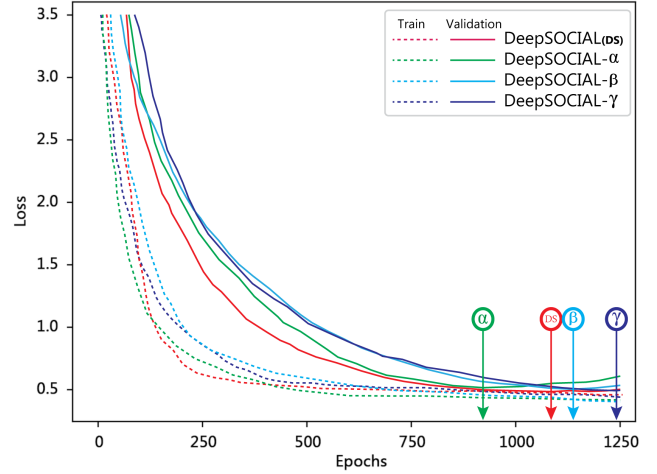


Figure 10. Training and Validation loss of the DeepSOCIAL models over the Open Images dataset.

Interestingly, the Faster-RCNN model shows good generalizability; however, its low speed is an issue which seems to be due to the computational cost of the “region proposal” technique. Since the system requires a real-time performance, any model with the speeds lower than 10 *fps* and/or a low level of accuracy may not be a suitable option for Social Distancing monitoring. Therefore, SSD and Faster-RCNN fail in this benchmarking assessment, despite their popularity in other applications. YOLOv3 and v4 provide relatively better results comparing the other models, and finally, the proposed DeepSOCIAL-DS model outperforms in terms of both speed and accuracy.

Figure 11 provides sample footage of the challenging scenarios when the people either enter or exit the scene, and only part of their body (e.g. their feet) are visible. The figure clearly indicates the strength of DeepSOCIAL in Row (a), comparing to the state-of-the-art. The bottom row (d) with blue bounding boxes shows the ground truth while some of the existing people with partial visibility are not annotated even in original ground truth dataset. Row (c), YOLOv3 shows a couple of more detections; however, the IoU of the suggested bounding boxes are low and some of them can be counted as False Positives. Row (b), the standard YOLOv4-based detector, shows a significant improvement comparing to row (c) and is considered as the second-best. Row (a), the DeepSOCIAL, shows 10 more true positive detections (highlighted by vertical arrows) comparing to the seconds best approach.

4.2 Social Distancing Evaluations

We considered the midpoint of the bottom edge of the detected bounding boxes as our reference points (i.e. shoes’ location). After the IPM, we would expect to have the location of each person, in the homogeneous space of BEV with a linear distance representation.

Any two people P_i, P_j with the Euclidean distance of smaller than r (i.e. the set restriction) in the BEV space

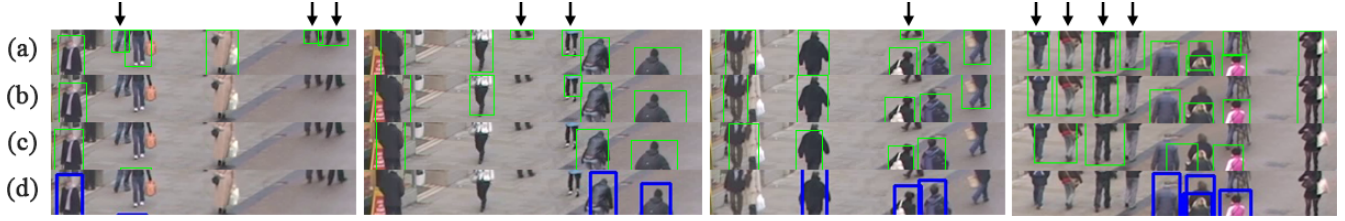


Figure 11. Human detection with partial visibility (missing upper body parts). (a) DeepSOCIAL (b) YOLOv4 trained on MS-COCO, (c) YOLOv3 (d) Ground-truth annotations from the OTC dataset.

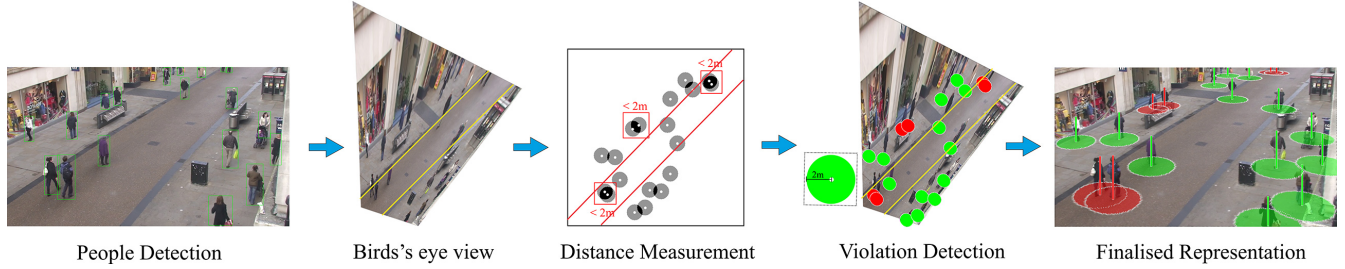


Figure 12. The summary of the steps taken in sections 3.1 to 3.3 for people detection, tracking, and distance estimation.

are considered as contributors in social distancing violation:

Depending on the type of overlapping and the violation assessment criteria, we define a violation detection function V with the input parameters of a pixel metrics ξ , the set safe distance of r (e.g 6ft or $\approx 2m$), the position of the query human H_q , and the closest surrounding person P_o .

$$V = \Lambda_{\xi}(H_q, P_o, r) \quad (24)$$

where ξ is number of pixels equal to 1.0 meter in the BEV space.

Figure 12.left (from Oxford Town Centre Dataset²⁹) shows the detected people followed by the steps we have taken for inter-people distance estimation including tracking, IPM, homogeneous 360° distance estimation, safe movements (people in green circles) and the violating people (with overlapping red circles):

$$\Lambda_{\xi}(P_i, P_j, r) = \begin{cases} 1 & \text{if } \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \leq r \\ 0 & \text{if } \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} > r \end{cases} \quad (25)$$

Regarding the Oxford Town Centre (OTC) dataset, every 10 pixels in the BEV space is equivalent to 98cm in the real world. Therefore, $r \approx 2 \times \xi$ and equal to 20 pixels. The inter-people distance measurement was measured based on the Euclidean $L2$ norm distance:

One of the controversial opinions that we received from health authorities was the way of dealing with family members and couples in social distancing monitoring. Some researchers believed social distancing should apply on every single individual without any exceptions and others were advising the couples and family members can walk in a close proximity

without being counted as breach of social distancing. In some countries such as in the UK and EU region the guideline allows two family members or a couple walk together without considering it as the breach of social distancing. We also considered a solution to activate the couple detection. This will be helpful when we aim at recognising risky zones based on the statistical analysis of overall movements and social distancing violations over a mid or long period (e.g. from few hours to few days).

Applying a temporal data analysis approach, we consider two individuals (p_i, p_j) as a couple, if they are less than d meters apart in an adjacency, for a t_{Λ} of more than ε seconds. As an example, in Figure 13(a), we have identified people who have been less a meter apart from each other for more than $\varepsilon = 5$ seconds, in the same moving trajectory:

$$\begin{aligned} C(p_i, p_j) = 1 & \quad \text{IF } p_i, p_j \in D \\ & \quad \text{AND } \Lambda_{\xi}(p_i, p_j, d) = 1 \\ & \quad \text{AND } t_{\Lambda}(p_i, p_j) > \varepsilon \end{aligned} \quad (26)$$

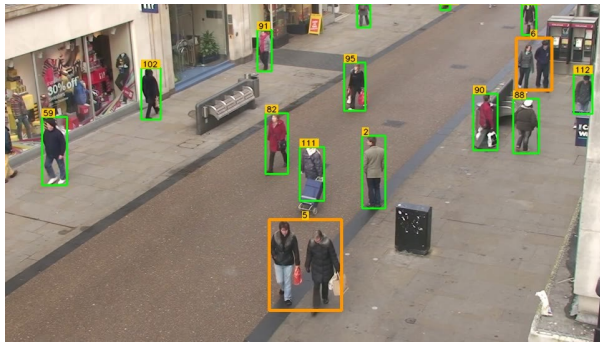
Figure 13(b) shows a sample of our representation for detected couples in a scene as well as multiple cases of social distancing violations. The yellow circles drawn for people diagnosed as couples have a radius of $(2 + \frac{d_c}{2})$ meter to ensure a minimum safety distance of 2m for each of them to their left and right neighbours. d_c is the distance of coupled members to each other.

In cases where a breach occurs between two neighbouring couple, or between a couple and an individual, all of the involved people will turn to red status, regardless of being a couple or not.

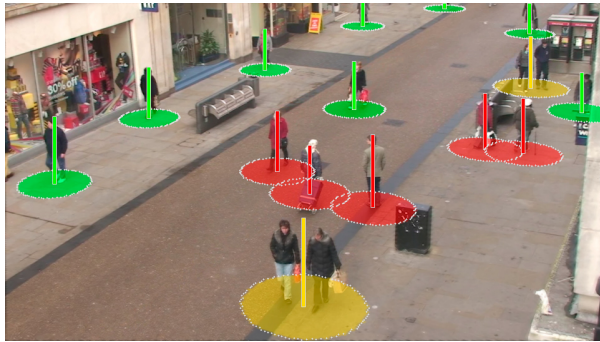
The flexibility of our algorithm in considering different types of scenarios enables the policymakers and health au-

Table 2. Speed, accuracy, and recall-rate comparison for seven DNN models on the Oxford Town Centre dataset.

Method	Backbone	Precision	Recall	FPS
DeepSOCIAL - α	CSPResNeXt50 - PANet - SPP - SAM	99.8%	96.7	23.8
DeepSOCIAL - β	CSPDarkNet53 - PANet - SPP	99.5%	97.1	24.1
DeepSOCIAL - γ	CSPResNeXt50 - PANet - SPP	99.6%	96.7	23.8
DeepSOCIAL (DS)	CSPDarkNet53 - PANet - SPP - SAM	99.8%	97.6	24.1
YOLOv3	DarkNet53	84.6%	68.2	23
SSD	VGG-16	69.1%	60.5	10
Faster R-CNN	ResNet-50	96.9%	83.0	3



(a) Examples of coupled people detections- orange bounding boxes



(b) Three types of detections: safe, violations, coupled

Figure 13. Social distancing violation detection for coupled people and individuals.

thorities to proceed with different types of investigations and evaluations for the spread of the infection.

For example, Figure 14 from the Oxford Town Centre dataset, provides a basic statistics about the number of people in each frame, the number of people who do not observe the distancing, the number of social distancing violations without counting the coupled groups as violations.

Regarding the coupled group we reached to an accuracy and recall rate of 98.7% and 23.9 *fps*, respectively which is slightly lower than our results of normal human detection as per the Table 2. This was expected due to added complexity in

tracking two side by side people and possibly more complex occlusion scenarios.

4.3 Zone-based Risk Assessment

We also tested the effectiveness of our model in assessing the long-term behaviour of the people. This can be valuable for health sector policy makers and governors to make timely decisions to save lives and reduce the consequent costs. Our experiments provided very interesting results that can be crucial to control the infection rates before it raises uncontrolled and unexpectedly.

In addition to people inter-distance measurement, we considered a long-term Spatio-temporal zone-based statistical analysis by tracking and logging the movement trajectory of people, density of each zone, total number of people who violated the social-distancing measures, the total time of the violations for each person and as the whole, identifying high-risk zones, and ultimately, creating an informative risk heat-map.

In order to perform the analysis, a 2-D grid matrix $G_t \in \mathbb{R}^{w \times h}$ (initially filled by zero) is created to keep the latest location of individuals using the input image sequences. G_t represents the status of the matrix at time t and w and h are the width and height of the input image I , respectively.

To consider environmental noise and better visualisation of the intended heat-map, every person was associated with a 3×3 Gaussian kernel k :

$$k(i, j) = \begin{bmatrix} 0_{i-1, j-1} & 1_{i, j-1} & 0_{i+1, j-1} \\ 1_{i-1, j} & 2_{i, j} & 1_{i+1, j} \\ 0_{i-1, j+1} & 1_{i, j+1} & 0_{i+1, j+1} \end{bmatrix} \quad (27)$$

where i and j indicate the centre point of the predicted box for each individual, p .

The grid matrix G will be updated for every new frame and accumulates the latest information of the detected people.

$$G_t = G_{t-1} + P_{(x,y)}^t \quad \forall P \in D_t \quad (28)$$

Figure 15 shows a sample representation of the accumulated tracking map after 500 frames of continuous people detection.

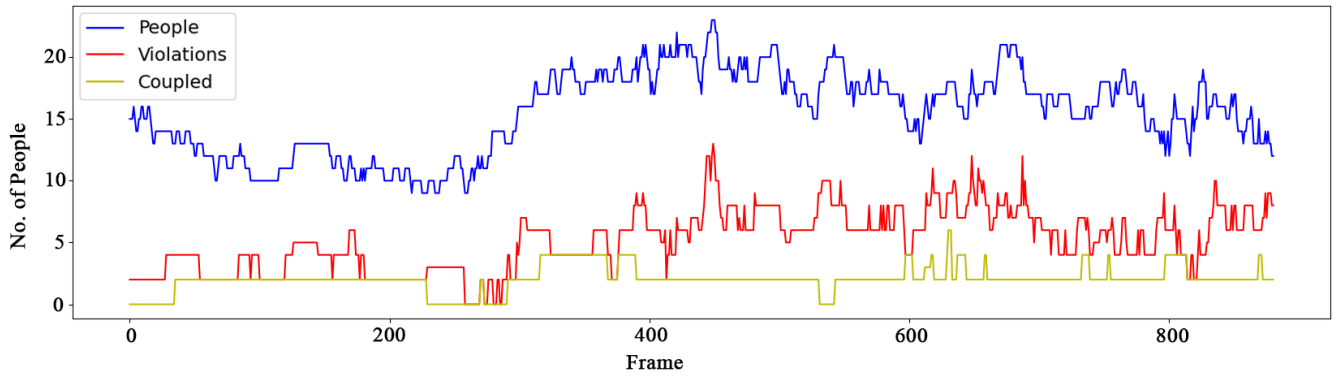


Figure 14. A 2D recording of the number of detected people in 900 frames from the OTC Dataset, as well as the number of violations and number of couples with no violations.

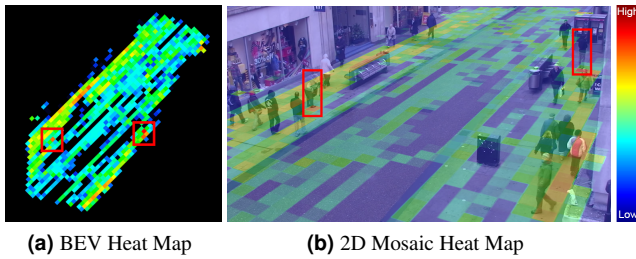


Figure 15. Accumulated tracking maps after 500 frames.

Since COVID-19 is an airborne virus, breathing by any static person in a fixed location can increase the density of the contamination on that point (assuming the person may carry the COVID-19), particularly in covered places with minimal ventilation. Therefore we can assign a more contamination weight to the steady state people.

Figure 15(b) and 15(a) shows two instances of cases where two people have been steady in two particular locations of the grid for a long period; hence, the heat map is turning to red for those locations. Both side-walks also show stronger heat map than the middle of the street due to the higher traffic of people movements. In general, the more red grids potentially indicates more risky spots.

In addition to people raw movement and tracking data, that would be more beneficial to analyse the density and the location of the people who particularly violated the social distancing measures.

In order to have a comprehensive set of information we aim to represent a long term heat-map of the environment based on the combination of accumulated detections, movements, steady states people, and total number of breaches of the social-distancing. This helps to identify risky zones, or redesign the layout of the environment to make it a safer place, or to apply more restrictions and/or limited access to particular zones. The restriction rules may vary depending on the application and the nature of the environment (e.g. this can vary in a hospital compared to a school).

Applying the social distancing violation criteria as per Eq. 24, we identify each individual in one of the following

categories:

- **Safe:** All people who observe the social distancing (green circles).

$$Z_g^t = \{P^t | V_P^t = 0\} \quad (29)$$

- **High-risk:** All people who violate the social distancing (red circles).

$$Z_r^t = \{P^t | V_P^t = 1\} \quad (30)$$

- **Potentially risky:** Those people who move together (yellow circles) and are identified as coupled. Any two people in a coupled group are considered as one identity as long as they do not breach the social distancing measures with their neighbouring people.

$$Z_y^t = \{(P^t | P \in C, P \notin Z_r)\} \quad (31)$$

We also use a 3-D violation matrix $S \in \mathbb{R}^{w \times h \times 3}$ to record the location of breaches for each person and its type (red or yellow):

$$S_{(x,y)}^t = S_{(x,y)}^{t-1} + \sum_{t=1}^n \left(R_{(x,y)}^t \cdot \alpha + T_{(x,y)}^t \cdot \beta + Y_{(x,y)}^t \cdot \delta \right) \quad (32)$$

$\forall R \in Z_r, T \in G_t, Y \in Z_y$

where R , T , and Y indicate cases with a violation, tracked people, and couples, respectively. α , β , and δ are the relative coefficients that can be set by health-related researches depending on the importance of each factor in spreading the virus.

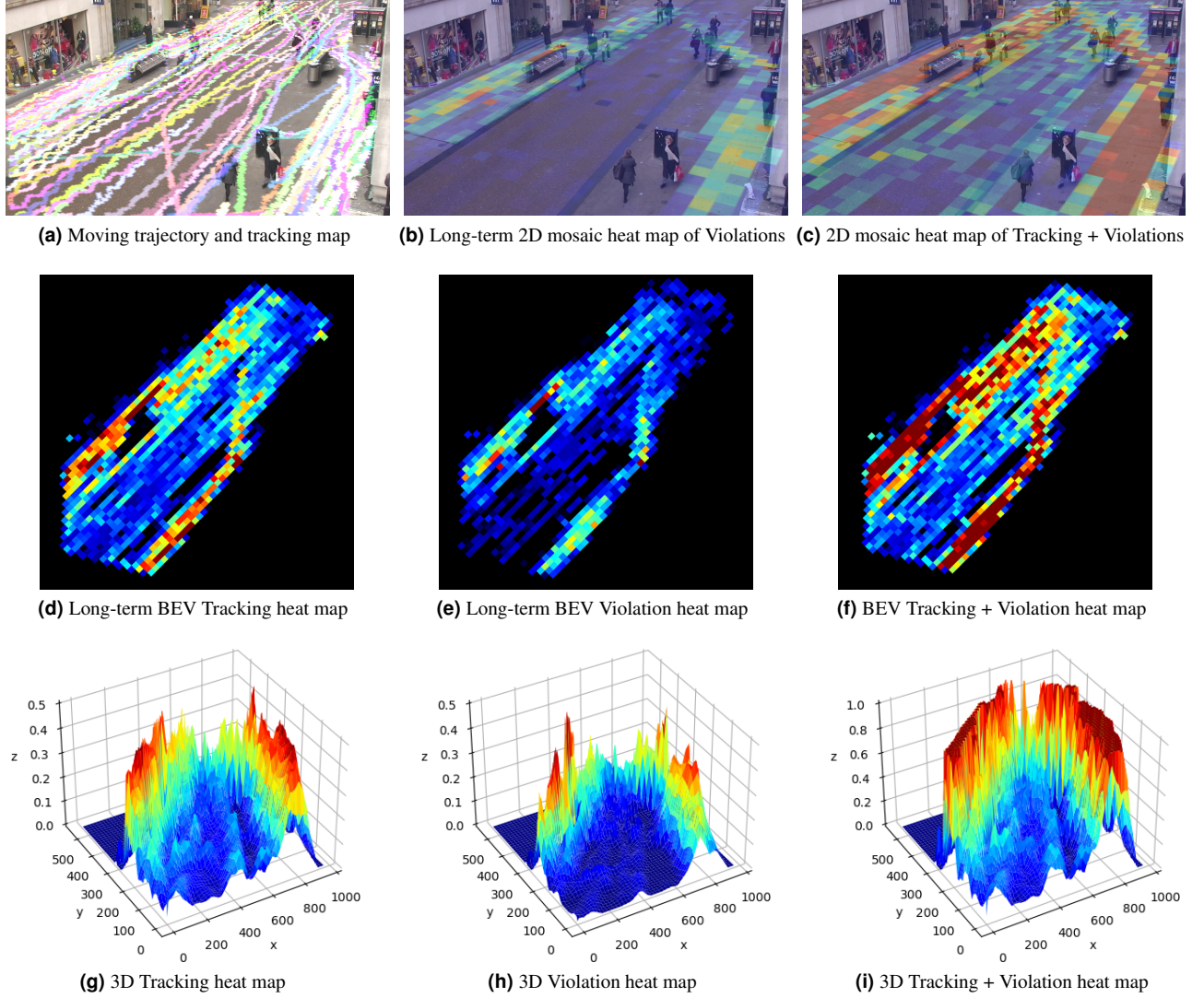


Figure 16. Data analysis based on people detections, tracking, movements, and breaches of social distancing measurements.

In order to visualise the 3D heat map of safe and red zones, we normalise the collected data in Eq. 32 as follows:

$$N(X, l, u) = l + (u - l) \times \frac{X - \min(X)}{\max(X) - \min(X)} \quad (33)$$

where X is the non-normalised values, l and u are the lower bound and upper bound of the normalisation matrix that we use to define the Hue colour range on the HSV channel.

Figure 16 shows a visualised output of the discussed experiments including tracking, 2-D and 3-D analytical heat-map of risky zones for the Oxford Town Centre dataset in HSV colour space:

$$R_t = N(\max(G^t, 2 \times S^t), 0, 120) \quad (34)$$

$$R \in \mathbb{R}^{w \times h}$$

Figure 16(a) shows the tracking paths of the passed people after 2500 frames. Figure 16(b) illustrates a mosaic heatmap

of people with social distancing violations only. Figure 16(c) shows the summed heatmap of social distancing violations and tracking. Figures 16(d), (e), and (f) show the birds-eye view heatmap of long term-tracking, violations, and the mixed heatmap, respectively. Figures 16(g), (h), and (i) are the corresponding 3D representation of the same heatmaps in the second row, for better visualisation of safe and risky zones.

The above configuration was an accumulating approach where all the violations and risky behaviours were added together in order to highlight potentially risky zones in a covered area with poor ventilation.

We also thought about cases when there exist a good chance of ventilation where the spread of virus would not be necessarily accumulative. In such cases we considered both increasing and decreasing counts depending on the overall time spent by each individual in each grid cell of the image, as well as the total absence time of individuals which potentially allows bring down the level of contamination.

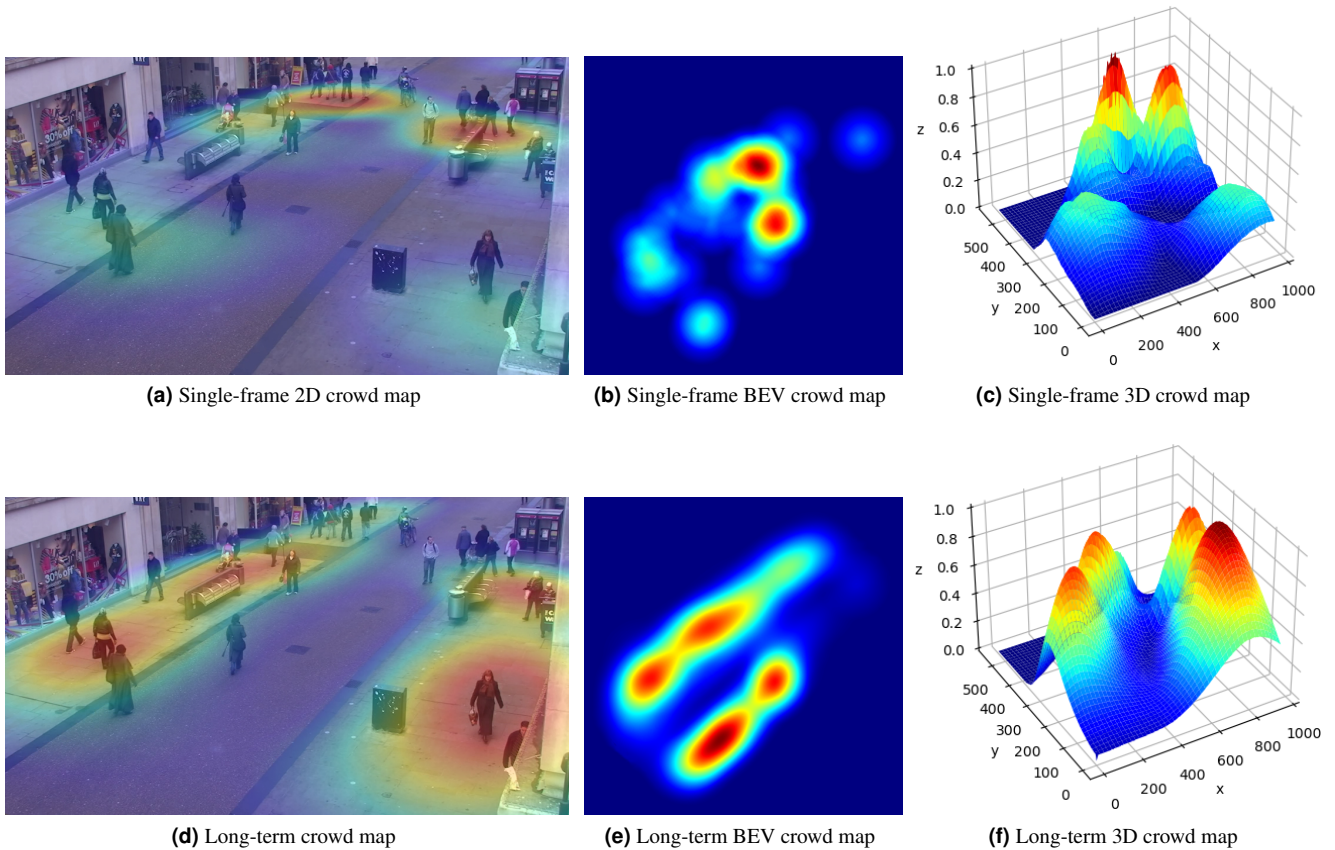


Figure 17. Single frame vs. Long-term crowd map (2D, BEV, 3D).

Figure 17 which we named it as crowd map shows the 2D and 3D representation of violations and risk heat maps where we applied both increasing and decreasing contamination trends. The first row of the image represents a single frame analysis with two peak zones. As can be seen, those zones belong to two crowded zones where two large groups of people are walking together and breach the social distancing rule. However, in other parts of the street a minimal level or risk is identified. This is due to large inter people distances and a consequent time gaps which allows breathing and therefore a decreasing rate of contamination.

The second row of Figure 17 shows a long-term crowd map which does not necessarily depend on the current frame. This can be a weighted averaging over all of the previous single-frame crowd maps.

One of the extra research questions in Figure 16 and 17 is how to define appropriate averaging weights and coefficients (α, β, γ) in Equation 32 and how to normalise the maps over the time. This is out of the scope of this research and needs further studies. However, here we aimed at showing the feasibility of considering a diversity of cases using the proposed method with a high level of confidence and accuracy in social distancing monitoring and risk assessment with the help of AI and Computer Vision.

5 Conclusion

We proposed a Deep Neural Network-Based human detector model called DeepSOCIAL to detect and track static and dynamic people in public places in order to monitor social distancing metrics in COVID-19 era and beyond. We utilised a CSPDarkNet53 backbone along with an SPP/PAN and SAM neck, Mish activation function, and Complete IoU loss function and developed an efficient and accurate human detector, applicable in various environments using any type of CCTV surveillance cameras. The system was able to perform in a variety of challenges including, occlusion, lighting variations, shades, and partial visibility. The proposed method was evaluated using large and comprehensive datasets and proved a major development in terms of accuracy and speed compared to three state-of-the-art techniques. The system performed real-time using a basic hardware and GPU platform.

We also conducted a zone based infection risk assessment and analysis to the benefit of the health authorities and governments. The outcome of this research is applicable for a wider community of researchers not only in computer vision, AI, and health sectors, but also in other industrial applications such as pedestrian detection in driver assistance systems, autonomous vehicles, anomaly behaviour detections, and variety of surveillance security systems.

References

1. World Health Organisation. WHO Corona-viruses Disease Dashboard (August 2020). Available at <https://covid19.who.int/table>.
2. WHO Director, Generals. Opening remarks at the media briefing on COVID-19 (2020). WHO generals and directors speeches.
3. Olsen, S. J. *et al.* Transmission of the severe acute respiratory syndrome on aircraft. *New Engl. J. Medicine* **349**, 2416–2422, DOI: [10.1056/NEJMoa031349](https://doi.org/10.1056/NEJMoa031349) (2003).
4. Adlhoch, C. *et al.* Considerations relating to social distancing measures in response to the COVID-19 epidemic (2020). European Centre for Disease Prevention and Control, Technical report.
5. Ferguson, N. M. *et al.* Strategies for mitigating an influenza pandemic. *Nature* **442**, 448–452, DOI: <https://doi.org/10.1038/nature04795> (2006).
6. Thu, T. P. B., Ngoc, P. N. H., Hai, N. M. *et al.* Effect of the social distancing measures on the spread of COVID-19 in 10 highly infected countries. *Sci. Total. Environ.* 140430, DOI: <https://doi.org/10.1016/j.scitotenv.2020.140430> (2020).
7. Morato, M. M., Bastos, S. B., Cajueiro, D. O. & Normey-Rico, J. E. An optimal predictive control strategy for COVID-19 (SARS-CoV-2) social distancing policies in Brazil. *Elsevier Annu. Rev. Control.* DOI: <https://doi.org/10.1016/j.arcontrol.2020.07.001> (2020).
8. Fong, M. W. *et al.* Nonpharmaceutical measures for pandemic influenza in nonhealthcare settings—social distancing measures. *Emerg. infectious diseases* **26**, 976, DOI: [10.3201/eid2605.190995](https://doi.org/10.3201/eid2605.190995) (2020).
9. F, A., N, Z. & A, U. Effectiveness of workplace social distancing measures in reducing influenza transmission: a systematic review. *BMC public health* 1–13, DOI: [10.1186/s12889-018-5446-1](https://doi.org/10.1186/s12889-018-5446-1) (2018).
10. Australian Government Department of Health. Deputy chief medical officer report on COVID-19. *Dep. Heal. Soc. distancing for coronavirus* DOI: <https://doi.org/10.1136/bmj.m1845> (2020).
11. Rezaei, M. & Shahidi, M. Zero-shot learning and its applications from autonomous vehicles to COVID-19 diagnosis: A review. *SSRN Mach. Learn. J.* **3**, 1–27, DOI: [10.2139/ssrn.3624379](https://doi.org/10.2139/ssrn.3624379) (2020).
12. Toğaçar, M., Ergen, B. & Cömert, Z. COVID-19 detection using deep learning models to exploit social mimic optimization and structured chest X-ray images using fuzzy color and stacking approaches. *Comput. Biol. Medicine* 103805, DOI: <https://doi.org/10.1016/j.compbimed.2020.103805> (2020).
13. Ulhaq, A., Khan, A., Gomes, D. & Paul, M. Computer vision for COVID-19 control: A survey. *Image Video Process.* DOI: [10.31224/osf.io/yt9sx](https://doi.org/10.31224/osf.io/yt9sx) (2020).
14. Nguyen, T. T. Artificial intelligence in the battle against coronavirus (COVID-19): a survey and future research directions. *arXiv Prepr.* **10**, DOI: [10.13140/RG.2.2.36491.23846/1](https://doi.org/10.13140/RG.2.2.36491.23846/1) (2020).
15. Choi, W. & Shim, E. Optimal strategies for vaccination and social distancing in a game-theoretic epidemiological model. *J. Theor. Biol.* 110422, DOI: <https://doi.org/10.1016/j.jtbi.2020.110422> (2020).
16. C, E., K, P. & S, W. J. Systematic biases in disease forecasting—the role of behavior change. *J. Epidemics* 96–105, DOI: [10.1016/j.epidem.2019.02.004](https://doi.org/10.1016/j.epidem.2019.02.004) (2019).
17. O, K. W. & G, M. A. A contributions to the mathematical theory of epidemics–i. *The Royal Soc. publishing* DOI: <https://doi.org/10.1098/rspa.1927.0118> (1991).
18. Heffernan, J. M., Smith, R. J. & Wahl, L. M. Perspectives on the basic reproductive ratio. *J. Royal Soc. Interface* **2**, 281–293, DOI: [10.1098/rsif.2005.0042](https://doi.org/10.1098/rsif.2005.0042) (2005).
19. Gupta, R., Pandey, G., Chaudhary, P. & Pal, S. K. Machine learning models for government to predict COVID-19 outbreak. *Int. J. Digit. Gov. Res. Pract.* **1**, DOI: [10.1145/3411761](https://doi.org/10.1145/3411761) (2020).
20. Nguyen, C. T. *et al.* Enabling and emerging technologies for social distancing: A comprehensive survey. *arXiv preprint* DOI: <https://arxiv.org/abs/2005.02816> (2020).
21. C, R. T. Game theory of social distancing in response to an epidemic. *PLoS computational biology* 1–9, DOI: [10.1371/journal.pcbi.1000793](https://doi.org/10.1371/journal.pcbi.1000793) (2010).
22. Ainslie, K. E. *et al.* Evidence of initial success for china exiting COVID-19 social distancing policy after achieving containment. *Wellcome Open Res.* **5**, DOI: <https://doi.org/10.12688/wellcomeopenres.15843.1> (2020).
23. Vidal-Alaball, J. *et al.* Telemedicine in the face of the COVID-19 pandemic. *Atencion primaria* **52**, 418–422, DOI: <https://doi.org/10.1016/j.aprim.2020.04.003> (2020).
24. Sonbhadra, S. K., Agarwal, S. & Nagabhushan, P. Target specific mining of COVID-19 scholarly articles using one-class approach. *J. Chaos, Solitons & Fractals* **140**, DOI: [10.1016/j.chaos.2020.110155](https://doi.org/10.1016/j.chaos.2020.110155) (2020).
25. Punn, N. S. & Agarwal, S. Automated diagnosis of COVID-19 with limited posteroanterior chest X-ray images using fine-tuned deep neural networks. *Image Video Process.* DOI: [arXiv:2004.11676](https://arxiv.org/abs/2004.11676) (2020).
26. Punn, N. S., Sonbhadra, S. K. & Agarwal, S. COVID-19 epidemic analysis using machine learning and deep learning algorithms. *medRxiv* DOI: <https://doi.org/10.1101/2020.04.08.20057679> (2020).
27. Jhunjhunwala, A. Role of telecom network to manage COVID-19 in india: Aarogya Setu. *Transactions Indian Natl. Acad. Eng.* 1–5, DOI: [10.1007/s41403-020-00109-7](https://doi.org/10.1007/s41403-020-00109-7) (2020).

28. Robakowska, M. *et al.* The use of drones during mass events. *Disaster Emerg. Medicine J.* **2**, 129–134, DOI: [10.5603/DEMJ.2017.0028](https://doi.org/10.5603/DEMJ.2017.0028) (2017).
29. Harvey, J., Adam. LaPlace. MegaPixels: Origins, ethics, and privacy implications of publicly available face recognition image datasets (2019).
30. Xin, T. *et al.* Freesense. *Proc. ACM on Interactive, Mobile, Wearable Ubiquitous Technol.* **2**, 1 – 23, DOI: <https://dl.acm.org/doi/10.1145/3264953> (2018).
31. Shi, F. *et al.* Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for COVID-19. *IEEE reviews biomedical engineering* DOI: [10.1109/RBME.2020.2987975](https://doi.org/10.1109/RBME.2020.2987975) (2020).
32. Hossain, F. A., Lover, A. A., Corey, G. A., Reigh, N. G. & T, R. FluSense: a contactless syndromic surveillance platform for influenzalike illness in hospital waiting areas. In *ACM Int. Joint Conference on Pervasive and Ubiquitous Computing*, 1–28, DOI: <https://doi.org/10.1145/3381014> (2020).
33. Polese, M. *et al.* Machine learning at the edge: A data-driven architecture with applications to 5G cellular networks. *IEEE Transactions on Mob. Comput.* DOI: [10.1109/TMC.2020.2999852](https://doi.org/10.1109/TMC.2020.2999852) (2020).
34. Brighente, A., Formaggio, F., Di Nunzio, G. M. & Tomasin, S. Machine learning for In-Region location verification in wireless networks. *IEEE J. on Sel. Areas Commun.* **37**, 2490–2502, DOI: [10.1109/JSAC.2019.2933970](https://doi.org/10.1109/JSAC.2019.2933970) (2019).
35. Liu, L. *et al.* Deep learning for generic object detection: A survey. *Int. journal computer vision* **128**, 261–318, DOI: <https://doi.org/10.1007/s11263-019-01247-4> (2020).
36. Rezaei, M., Sarshar, M. & Sanaatiyan, M. M. Toward next generation of driver assistance systems: A multimodal sensor-based platform. In *2010 The 2nd International Conference on Computer and Automation Engineering (ICCAE)*, vol. 4, 62–67 (IEEE, 2010).
37. Sabzevari, R., Shahri, A., Fasih, A., Masoumzadeh, S. & Ghahroudi, M. R. Object detection and localization system based on neural networks for robo-pong. In *2008 5th International Symposium on Mechatronics and Its Applications*, 1–6 (IEEE, 2008).
38. Nguyen, D. T., Li, W. & Ogunbona, P. O. Human detection from images and videos: A survey. *Int. J. Pattern Recognit.* **51**, 148–175, DOI: <https://doi.org/10.1016/j.patcog.2015.08.027> (2016).
39. Serpush, F. & Rezaei, M. Complex human action recognition in live videos using hybrid FR-DL method. *arXiv preprint, Comput. Vis. Pattern Recognit.* DOI: [arXiv:2007.02811](https://arxiv.org/abs/2007.02811) (2020).
40. Gawande, U., Hajari, K. & Golhar, Y. Pedestrian detection and tracking in video surveillance system: Issues, comprehensive review, and challenges. *Recent Trends Comput. Intell. Intech Open Publ.* DOI: [10.5772/intechopen.90810](https://doi.org/10.5772/intechopen.90810) (2020).
41. Punn, N. S., Sonbhadra, S. K. & Agarwal, S. Monitoring COVID-19 social distancing with person detection and tracking via fine-tuned YOLO v3 and deepsort techniques. *arXiv Prepr.* DOI: [arXiv:2005.01385](https://arxiv.org/abs/2005.01385) (2020).
42. Wojke, N., Bewley, A. & Paulus, D. Simple online and realtime tracking with a deep association metric. In *IEEE International Conference on Image Processing (ICIP)*, 3645–3649, DOI: [10.1109/ICIP.2017.8296962](https://doi.org/10.1109/ICIP.2017.8296962) (2017).
43. Khandelwal, P. *et al.* Using computer vision to enhance safety of workforce in manufacturing in a post COVID world. *Comput. Vis. Pattern Recognit.* DOI: [10.1016/S0140-6736\(20\)30185-9](https://doi.org/10.1016/S0140-6736(20)30185-9) (2020).
44. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L. MobileNetV2: Inverted residuals and linear bottlenecks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, DOI: [10.1109/CVPR.2018.00474](https://doi.org/10.1109/CVPR.2018.00474) (2018).
45. Yang, D., Yurtsever, E., Renganathan, V., Redmill, K. & Özgüner, U. A vision-based social distancing and critical density detection system for COVID-19. *Image video Process.* DOI: [arXiv:2007.03578](https://arxiv.org/abs/2007.03578) (2020).
46. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, DOI: [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2015).
47. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *IEEE International Conference on Computer Vision and pattern Recognition (CVPR)*, DOI: [10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90) (2015). [1512.03385](https://arxiv.org/abs/1512.03385).
48. Wang, C.-Y. *et al.* CSPNet: A new backbone that can enhance learning capability of CNN. In *IEEE CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 390–391, DOI: [10.1109/CVPRW50498.2020.00203](https://doi.org/10.1109/CVPRW50498.2020.00203) (2020).
49. Tan, M. & Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Thirty-sixth International Conference on Machine Learning (ICML)*, DOI: <https://doi.org/10.1101/2020.08.13.20173997> (2019).
50. Girshick, R., Donahue, J., Darrell, T. & Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 580–587, DOI: [10.1109/CVPR.2014.81](https://doi.org/10.1109/CVPR.2014.81) (2014).
51. Girshick, R. Fast R-CNN. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, 1440–1448, DOI: [10.1109/ICCV.2015.169](https://doi.org/10.1109/ICCV.2015.169) (2015).
52. Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis Mach.*

- Intell. (T-PAMI)* **39**, 1137–1149, DOI: [10.1109/TPAMI.2016.2577031](https://doi.org/10.1109/TPAMI.2016.2577031) (2017).
53. Sharifi, A., Zibaei, A. & Rezaei, M. DeepHAZMAT: Hazardous materials sign detection and segmentation with restricted computational resources. *Eng. Res.* DOI: <http://dx.doi.org/10.2139/ssrn.3649600> (2020).
 54. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You Only Look Once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779–788, DOI: [10.1109/CVPR.2016.91](https://doi.org/10.1109/CVPR.2016.91) (2016).
 55. Redmon, J. & Farhadi, A. YOLO9000: Better, faster, stronger. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6517–6525, DOI: [10.1109/CVPR.2017.690](https://doi.org/10.1109/CVPR.2017.690) (2017).
 56. Redmon, J. & Farhadi, A. YOLOv3: An incremental improvement. *Comput. Vis. Pattern Recognit. (cs.CV)* DOI: [arXiv:1804.02767](https://arxiv.org/abs/1804.02767) (2018).
 57. Bochkovskiy, A., Wang, C.-Y. & Liao, H.-Y. M. YOLOv4: Optimal speed and accuracy of object detection. *arXiv Comput. Vis. Patter Recognit.* DOI: [arXiv:2004.10934](https://arxiv.org/abs/2004.10934) (2020).
 58. Liu, W. *et al.* SSD: Single shot multibox detector. *Eur. Conf. on Comput. Vis. (ECCV)* 21—37, DOI: [10.1007/978-3-319-46448-0_2](https://doi.org/10.1007/978-3-319-46448-0_2) (2016).
 59. Everingham, M., Van Gool, L., Williams, C., Winn, J. & Zisserman, A. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. *Int. J. Comput. Vis. (IJCV)* **88**, 303–338, DOI: <https://doi.org/10.1007/s11263-009-0275-4> (2010).
 60. Chen, X. *et al.* Microsoft COCO captions: Data collection and evaluation server. *arXiv Comput. Vis. Pattern Recognit.* DOI: [arXiv:1504.00325](https://arxiv.org/abs/1504.00325) (2015).
 61. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261–2269, DOI: [10.1109/CVPR.2017.243](https://doi.org/10.1109/CVPR.2017.243) (2017).
 62. Lin, T.-Y. *et al.* Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125, DOI: [10.1109/CVPR.2019.00656](https://doi.org/10.1109/CVPR.2019.00656) (2017).
 63. Liu, S., Qi, L., Qin, H., Shi, J. & Jia, J. Path aggregation network for instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 8759–8768, DOI: [10.1109/CVPR.2018.00913](https://doi.org/10.1109/CVPR.2018.00913) (2018).
 64. Woo, S., Park, J., Lee, J.-Y. & Kweon, I. S. CBAM: Convolutional block attention module. In *European Conference on Computer Vision (ECCV)*, 4–19, DOI: https://doi.org/10.1007/978-3-030-01234-2_1 (2018).
 65. He, K., Zhang, X., Ren, S. & Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *European Conference on Computer Vision (ECCV)*, 346–361, DOI: [10.1007/978-3-319-10578-9_23](https://doi.org/10.1007/978-3-319-10578-9_23) (Springer International Publishing, 2014).
 66. Zheng, Z. *et al.* Distance-IoU Loss: Faster and better learning for bounding box regression. In *AAAI Conference on Artificial Intelligence*, 1–8, DOI: [10.1609/aaai.v34i07.6999](https://doi.org/10.1609/aaai.v34i07.6999) (2019).
 67. Misra, D. Mish: A self regularized non-monotonic neural activation function. *Neural Evol. Comput.* DOI: [arXiv:1908.08681](https://arxiv.org/abs/1908.08681) (2019).
 68. Yao, Z., Cao, Y., Zheng, S., Huang, G. & Lin, S. Cross-iteration batch normalization. *Mach. Learn.* DOI: [arXiv:2002.05712](https://arxiv.org/abs/2002.05712) (2020).
 69. Ghiasi, G., Lin, T.-Y. & Le, Q. V. DropBlock: A regularization method for convolutional networks. In *Advances in Neural Information Processing Systems (NIPS)* (2018). [1810.12890](https://arxiv.org/abs/1810.12890).
 70. Müller, R., Kornblith, S. & Hinton, G. E. When does label smoothing help? In *Advances in Neural Information Processing Systems (NIPS)*, 4694–4703, DOI: [arXiv:1906.02629](https://arxiv.org/abs/1906.02629) (2019).
 71. Loshchilov, I. & Hutter, F. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 1–16, DOI: [arXiv:1608.03983](https://arxiv.org/abs/1608.03983) (2016).
 72. Kuznetsova, A. *et al.* The open images dataset v4. *Int. J. Comput. Vis.* 1–26, DOI: <https://doi.org/10.1007/s11263-020-01316-z> (2020).
 73. Bewley, A., Ge, Z., Ott, L., Ramos, F. & Upcroft, B. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, 3464–3468, DOI: [10.1109/ICIP.2016.7533003](https://doi.org/10.1109/ICIP.2016.7533003) (IEEE, 2016).
 74. Rezaei, M. & Klette, R. Computer vision for driver assistance. *Book, Springer Int. Publ.* DOI: [10.1007/978-3-319-50551-0](https://doi.org/10.1007/978-3-319-50551-0) (2017).
 75. Saleem, N. H., Chien, H.-J., Rezaei, M. & Klette, R. Effects of ground manifold modeling on the accuracy of stixel calculations. *IEEE Transactions on Intell. Transp. Syst.* **20**, 3675–3687 (2018).
 76. Russakovsky, O. *et al.* ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis. (IJCV)* **115**, 211–252, DOI: [10.1007/s11263-015-0816-y](https://doi.org/10.1007/s11263-015-0816-y) (2015).