Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Deep visual social distancing monitoring to combat COVID-19: A comprehensive survey

Yassine Himeur a,*, Somaya Al-Maadeed a, Noor Almaadeed b, Khalid Abualsaud b, Amr Mohamed c, Tamer Khattab c, Omar Elharrouss a

a Computer Science and Engineering Department, Qatar University, Qatar
b Electrical Engineering Department, Qatar University, Qatar

table of contents

1. Introduction

1.1. Preliminary

In December 2019, China officially announced the discovery of a new coronavirus disease, namely COVID-19, which is mainly caused by the severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2), which has been a source of a global epidemic (Ghaemi, Amir, Bahjii, Yuhana, & Ferrara, 2021). Until January 12, 2022, there have been 312,173,462 confirmed cases of COVID-19, including 5,501,000 deaths, reported to the world health organization (WHO) (World Health Organization, 2021). At that end, an increasing effort has been made by the research community to put in use intelligent tools and measures to reduce or slow down the spread of COVID-19. In this respect, various studies have investigated some of the main pandemic open challenges, such as those related to (i) predicting of COVID-19 risk in public environments using IoT and machine learning (ML) (Elbasi, Topcu, & Mathew, 2021; Ramchandani, Fan, & Mostafavi, 2020; Tang, Peng, Chiheb, & Fan, 2021), (ii) monitoring social distancing (SD) and detecting violations (Ar et al., 2020; Prabakaran, Kumar, Kiran, & Supriya, 2022), (iii) detecting whether people are wearing masks and if they are wearing them correctly (Qin & Li, 2020; Tomás, Rego, Viciano-Tudela, & Lloret, 2021), and (iv) processing of thermal imaging to detect COVID-19 (Teboubi, Messaoud, Hajjaji, & Mtibaa, 2021). After three COVID-19 waves, the growing number of new infections still reminds us of the importance of taking precautionary measures. SD and wearing masks have been proven to be efficient nonpharmaceutical intervention measures (Özbek, Syed, & Öksüz, 2021). They are low-cost, convenient, and noninvasive to slow the spread of COVID-19 and flatten the curves of infection (Srivastava, Zhao, Manay, & Chen, 2021). The efficacy of these measures is outstanding in large cities, where contact and interaction between people are expected in daily activities (work, travel, education, etc.). To that end, these measures have been considered mandatory practice by almost all countries. However, the failure to follow these procedures, the lack of timely prevention, and non-compliance to proper wearing of face masks can lead to higher infection rates. Therefore, developing effective methods to automatically detect SD violations, identify proper mask-wearing, and measure

* Correspondence to: Computer Science and Engineering, Qatar University, Qatar, Doha 2713, Qatar.
E-mail addresses: yassine.himeur@qu.edu.qa (Y. Himeur), s_alali@qu.edu.qa (S. Al-Maadeed), n.alali@qu.edu.qa (N. Almaadeed), k.abualsaud@qu.edu.qa (K. Abualsaud), anrm@qu.edu.qa (A. Mohamed), tkhattab@ieee.org (T. Khattab), elharrouss.omar@gmail.com (O. Elharrouss).

https://doi.org/10.1016/j.scs.2022.104064
Received 17 April 2022; Received in revised form 7 July 2022; Accepted 12 July 2022
Available online 21 July 2022
2210-6707/© 2022 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
Fig. 1. Smart DL-based VSDM system for smart cities: the most important steps are explained including (i) data collection, (ii) data storage, (iii) pedestrian detection, (iv) distance measurement and (vi) violation detection.

body temperatures have attracted significant attention (Farooqi & Usman, 2021; Gad, ElBary, Alkhedher, & Ghazal, 2020). Indeed, these works can provide the public with information about where the risk of COVID-19 transmission may be high.

SD plays a major role in slowing down the distribution of the COVID-19 virus. Although the distance to be preserved between people is country-specific, most of the studies have defined SD as maintaining at least a distance of two meters (six feet) apart from other persons to prevent potential contacts (Agarwal et al., 2021; Kumar, John, Vignesh, & Jagannath, 2022). Typically, while WHO has recommended one meter of physical distance, as adopted in France, Singapore, Hong Kong, Denmark and China, India, UK, Qatar and many other countries have been maintaining 2 meters’ distance. The importance of SD also comes from its substantial economic benefits as it has long-run recovery effects on economic development (Pooranam, Sushma, Sruthi, & Sri, 2021). The COVID-19 pandemic may not be ended in the near future, and automated systems with the ability to monitor and analyze whether people are respecting or not SD norms help significantly profit our society. Besides, recent improvements in ML and DL allow object detection techniques to be quite efficient, which has benefited researchers to measure and monitor SD behaviors among pedestrians in public areas by analyzing recorded videos from fixed surveillance (e.g., CCTV cameras) or drone-based surveillance (Elharrouss et al., 2021; Haq, Du, & Jan, 2022). Typically, vision-based IoT systems already installed in public areas can be augmented with the people detection capability, which is a sub-task of the generic object detection process (Gaisie, Oppong-Yeboah, & Cobbinah, 2022; Manzira, Charly, & Caulfield, 2022). Moving on, adequate measures can be then initiated to measure the physical distances between detected pedestrians. Fig. 1 illustrates an overall architecture of a VSDM system based on DL for smart cities applications.

Because monitoring, managing, and preventing the spread of the COVID-19 virus require innovative and intelligent solutions and path-breaking tools, ML models, and more particularly deep learning (DL) models, play a crucial role in the humanity’s battle during the pandemic. Typically, computer vision (CV), which is part of artificial intelligence (AI), can teach computers to comprehend visual scenes and analyze dense crowds (Mohamed & Abdel Samee, 2022). In this regard, machines have become able to (i) identify and track objects, (ii) measure distance between them, and (iii) respond to observed scenes using cameras, smartphones, and DL tools (Nagrath et al., 2021). Similarly, CV combined with DL has recently been used to capture the average amount of human activity, monitor SD behaviors, and detect the violation of face mask-wearing in major cities. Typically, face mask detection is considered a complementary task to SD monitoring to decrease the risk of contamination with COVID-19. Drones or unmanned aerial vehicles (UAVs) have also been utilized to fight the COVID-19 virus in open areas (e.g., the perimeters of sports facilities and stadiums) (Conte et al., 2021) by collecting biomedical data of individuals, (ii) monitoring SD and recording essential signs parameters (e.g., respiratory rates, body temperature, heart rates, etc.). This has been efficient for analyzing individuals' health status and limiting the spread of the virus.

Pedestrian detection is the most critical task in VSDM systems, and the efficacy of the SD analysis depends mainly on accurately detecting the pedestrians before measuring the distance between them. To that end, a great effort has been devoted to developing efficient convolutional neural networks (CNN)-based pedestrian detectors. Because pedestrian detectors are a particular type of object detectors that focus only on detecting pedestrians in images or video frames, it was rational to use already existing object detection schemes. Typically, neural network-based detectors have been extensively used by the research community to develop VSDM systems, including (i) single-stage object detectors such as you only look once (YOLO) (YOLOv1 (Redmon, Divvala, Girshick, & Farhadi, 2016), YOLOv2 (Redmon & Farhadi, 2017), YOLOv3 (Redmon & Farhadi, 2018), YOLOv4 and Wang, Bochkovskiy, and Liao (2021)), single shot multibox detector (SSD) (Liu et al., 2016); (ii) and two-stage detectors, such as region proposals (RCNN (Girshick, Donahue, Darrell, & Malik, 2014), Fast-RCNN (Girshick, 2015) Faster-RCNN (Ren, He, Girshick, & Sun, 2015), cascade-RCNN (Pang et al., 2019), mask-RCNN (He, Gkioxari, Dollár, & Girshick, 2017), etc.),
Retina-Net (Lin, Goyal, Girshick, He, & Dollár, 2017), single-shot refinement neural network for object detection (RefineDet) (Zhang, Wen, Bian, Lei, & Li, 2018), and deformable convolutional networks (Dai et al., 2017; Zhu, Hu, Lin, & Dai, 2019).

This review is introduced to shed light on the progress made by the scientific community in developing DL-based tools for VSDM since the pandemic’s start. Specifically, a well-designed taxonomy is introduced to better overview existing frameworks from various perspectives, including the surveillance type (i.e., fixed or mobile), methodology (hand-crafted-based or CNN-based), nature of pedestrian detectors (single-stage or two-stage), and complexity of CNN models (i.e., complex or light-weight), etc. Moreover, a comparative study is conducted to assess the competency of DL-based VSDM solutions, primarily based on CNN models. Thereafter, insightful observations are made to identify solved challenges and those that remain unresolved, such as pedestrian overlapping, real-time implementation, camera calibration, lack of annotated datasets, security and privacy concerns, etc. Additionally, future directions that can help improve the performance of VSDM and promote its implementation are highlighted. Overall, the main contributions of this paper can be summarized as follows:

- Presenting, to the best knowledge of the authors’ knowledge, the first review of deep VSDM literatures.
- Presenting the background of the VSDM concept and explaining its main steps.
- Summarizing datasets used for validating VSDM frameworks and discussing their characteristics and limitations.
- Systematically reviewing existing DL-based VSDM techniques and identifying their pros and cons.
- Analyzing and discussing the performance of existing DL-based VSDM solutions and presenting a comparative study of relevant works.
- Highlighting the open issues where the actual research effort is heading and providing insights about the future directions that can attract considerable interest in the near future.

1.2. Survey methodology

VSDM literatures have been surveyed by searching them in academic databases, including Scopus, Elsevier, IEEEXplore, Springer, WebolScience, etc. In doing so, the following keywords have been considered: “visual social distancing monitoring”, “social distancing detection using deep learning”, “social distancing analysis using computer vision” “social distancing monitoring using CNN” with the “document title, abstract and keywords” have been set in the advanced search. Hundreds of peer-reviewed articles have been obtained, but not all of them were related to the topic of the review. To that end, a careful filtering process has been conducted as follows: (i) all related journal papers have been included in this review as they present a detailed analysis and description, (ii) conferences papers written in other languages than English and not presenting sufficient quantitative results and experiments are filtered out, (iii) conferences lacking visual detection results have been excluded, and (iv) studies validated on small image datasets have not been considered. Moreover, some studies present very similar approaches, and only the datasets used to validate them are different. In this regard, only the frameworks validated on sufficient benchmarking data with a well-defined validation process have been included in this review. Overall, more than 75 VSDM literatures have been considered, covering peer-reviewed journal articles, conference proceedings articles, book chapters, and preprints.

The rest of this paper is organized as follows. Section 2 provides the background of VSDM systems, where the overall methodology is explained, and the types of adopted surveillance are described. Section 3 summarizes existing datasets used to validate VSDM techniques. Moving on, the limitations and drawbacks of non-visual SD monitoring (NVSDM) frameworks are briefly discussed in Section 4. Next, a thorough overview conducted based on a well-defined taxonomy of VSDM studies is presented in Section 5. After that, the important findings following this comprehensive review are identified in Section 6, where critical analysis is performed, and open challenges are highlighted. Lastly, future directions are derived in Section 7 before concluding this paper in Section 8.

2. Background

VSDM systems are based on detecting pedestrians, measuring the distance between them, and then quantifying the risk level of contaminating COVID-19 between monitoring people. Fig. 2 explains how the risk level can vary when monitored pedestrians are close to each other. Specifically, the closer or denser a crowd is, the more risky it is considered. The least risky level is on the top left while the more risky is on the top right.

Usually, a scene is defined at time t as a three-tuple \( S = (A, P_0, d_T) \), where \( A \in R^{H \times W \times 3} \) refers to the RGB frame. While \( H \) and \( W \) represent the height and width, receptively.

\( P_0 \in R \) represents the area of interest (ROI) on the real-world ground plane, and \( d_T \) stands for the physical distance threshold which is required to maintain a safe environment.

2.1. Problem definition

Given \( S \), VSDM techniques focus on detecting a list of individuals pose vectors \( A = (a_1, a_2, \ldots, a_n) \), \( a \in R^2 \) in the coordinates on the real-world ground plane along with the related list of interpersonal distances \( D = (d_1, d_2, \ldots, d_n) \in R^n \). \( n \) represents the number of individuals in the ROI.

2.2. Image to world mapping

At this stage, images are mapped into real-world, where the second mapping function \( h : a \rightarrow a' \) is obtained. Typically, \( h \) represents an inverse perspective transformation function, which enables mapping \( a \) in image coordinates to \( a' \) is obtained. Typically, \( h \) is represented in 2D bird’s eye view (BEV) coordinates, where the ground plane \( z = 0 \) is assumed. Specifically, the inverse homography transformation (Forsyth & Ponce, 2011) can be used to perform this:

\[
{a'}^h = M^{-1}a^w
\]

where \( M \in R^{3 \times 3} \) represents a transformation matrix that describes the translation and rotation from world to image coordinates. In this respect, \( a^w = [a_i', a_j'] \) refers to the homogenous representation of \( a' = [a_i, a_j] \) in image coordinates, and \( a^h = [a_i^h, a_j^h, 1] \) constitutes the homogeneous representation of the mapped pose vector. Moving on, the real-world pose vector \( a \) can be obtained from \( a^h \) with a = \([a_i^h, a_j^h, 1] \). This operation is essential since it facilitates the measurement of the real physical distances between each pedestrian pair.

2.3. Pedestrian detection

First, any VSDM system aims to detect individuals (or pedestrians) in the video frames collected using fixed or drone-based monocular or stereo cameras and insert a collection of bounding boxes (BBs) (Li, Varble, Turkbye, Xu, & Wood, 2022). Typically, an ML-based object detector \( O \) is applied on the frame \( a \):

\[
[T]_a = O(a)
\]

\( O : a \rightarrow [T]_a \) maps a frame \( a \) into \( n \) tuples \( T_i = (l_i, b_i, t_i) \), \( \forall i \in \{1, 2, \ldots, n\} \), where \( n \) represents the number of detected objects, \( l_i \in L \) is the object class label among the overall object label set \( L \), \( b_i = (b_{i,0}, b_{i,1}, b_{i,2}, b_{i,3}) \) represent the corresponding BB with four corners.
\( b_{i,j} = (x_{i,j}; y_{i,j}) \) provides pixel indices in the image domain, where \( j \) represents the corners at “top-left”, “top-right”, “bottom-left”, and “bottom-right”, respectively. Lastly, \( s_i \) indicates the corresponding detection score. VSDM systems attempt to only detect the case of \( l = \) “person”.

### 2.4. Social distancing (SD) detection

After detecting all the BBs \( B = (b_1, b_2, \ldots, b_n) \) in real coordinates, the corresponding list of interpersonal distances \( D \) is calculated between their centroids. Typically, the distance \( d_{x,y} \) for the individuals detected by the BBs \( b_1 \) and \( b_2 \) is estimated using the Euclidean distance between their centroids \( b'_1 \) and \( b'_2 \):

\[
d_{x,y} = \| b'_1 - b'_2 \| \quad (3)
\]

The overall number of SD violations \( v \) in a scene is computed as follows:

\[
V = \sum_{i=1}^{n} \sum_{j=1}^{n} v(d_{x,y}) \mid v(d_{x,y}) = 1 \text{ if } d_{x,y} < d_T \text{, otherwise } v(d_{x,y}) = 0 \quad (4)
\]

where \( d_T \) is the interpersonal distance. It is worth noting that the obtained violations can be filtered by imposing thresholds on the time contact patterns and/or the number of contacts, and (ii) considering family/non-family classification. For example, in Pouw, Toschi, van Schadewijk, and Corbetta (2020), a minimum contact time threshold \( \alpha = 5s \) is defined to tag SD offenders.

### 2.5. Tracking and reporting

Once two pedestrians are detected to be close to each other and the distance value violates the minimum SD norm, the color of the bounding box is updated/changed to red. Moreover, the BB information is saved in a violation database and transmitted to surveillance and monitoring center for reporting purposes and sending alarms to concerned offenders. On the hand, centroid tracking algorithms can be deployed to track the people violating/breaching the SD norm. For instance, Yang, Sun, et al. (2021) use the simple online and real-time tracking (SORT) algorithm (Bewley, Ge, Ott, Ramos, & Upcroft, 2016) to track pedestrians detected with YOLOv4 due to its simplicity and quick inference. Similarly, DeepSort (Wojke, Bewley, & Paulus, 2017), which is one of the most widely tracking algorithms is utilized in Punn, Sonbhadra, Agarwal, and Rai (2020) to track pedestrians detected with YOLOv3. The tracking has been performed using BBs and assigned IDs of people violating the SD norm. Other variants of DeepSort can be also utilized such as, StrongSort (Du, Song, Yang, & Zhao, 2022). Moreover, multi-object tracking (MOT) algorithms have also been considered to track detected pedestrians. This is the case of Al-Sa’d et al. (2022), where the global nearest neighbor (GNN) tracking technique has been used.

Fig. 3 explains the main steps of performing VSDM based on CNN.

### 2.6. Evaluation metrics

To quantify the performance of existing VSDM frameworks and inform the state-of-the-art, we perform a comparative analysis showing their original results on their own datasets. Accordingly, we first briefly present the evaluation metrics commonly used in VSDM studies, including accuracy, F1 score, average precision (AP), and mean average precision (mAP).

#### Accuracy:

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)
\]

#### F1 score:

\[
F1 = 2 \times \frac{\text{recall} \times \text{precision}}{\text{recall} + \text{precision}} \quad (6)
\]

where: \( \text{recall} = \frac{TP}{TP + FN} \) and \( \text{precision} = \frac{TP}{TP + FP} \). Additionally, \( TP \) and \( TN \) represent the true positives, and true negatives, respectively. While, \( FP \) and \( FN \) refer to false positives, and false negatives, respectively.

#### Mean average precision (mAP):

\[
mAP = \frac{1}{h} \sum_{h=1}^{K} AP_h \quad (7)
\]
where \( AP_k \) refers to the average precision of class \( k \). Overall, AP is defined as:

\[
AP_{\text{class}} = \int_0^1 P(r) \, dr
\]  

where class refers to the object classes, e.g., "pedestrian" and "non-pedestrian" or "people respecting SD" and "people violating SD", etc.

**Intersection over union (IoU):** when pedestrians are detected, the model can generate multiple BBs for a single pedestrian. Thus, the intersection over union (IoU)-based filter is used, which is calculated for areas of two BBs \( B_1 \) and \( B_2 \) as follows:

\[
IoU = \frac{B_1 \cap B_2}{B_1 \cup B_2}
\]  

IoU provides the similarity rate between the ground truth BB and the predicted BB as a measurement for the quality of the prediction, the value of IoU varies from 0 to 1.

### 2.7. Surveillance methodology

#### 2.7.1. Fixed surveillance

As the IP surveillance industry enters the era of AI, security network cameras (IP cameras) and closed-circuit television (CCTV) cameras have seen significant advances through applying AI and deep learning technologies. These next-generation cameras have been equipped with video analytics and high-performance computing power, allowing users to convert real-time video frames into the big-data analysis. VSDM based on fixed surveillance refers to using existing CCTV cameras and/or IP cameras combined with ML and computer vision capabilities for detecting if pedestrians are respecting the SD norms or not (Pandiyar et al., 2022). When two or more pedestrians are detected in close contact using object detectors and distance measurement algorithms, an alarm is produced to alert people found in the monitored environment. AI alerts are also sent to concerned authorities or guards who can ask people to maintain distance. Fixed surveillance is mainly used in indoor environments, such as shopping areas, airports, sports facilities (e.g., stadiums), etc. (Al-Sa’d et al., 2022).

#### 2.7.2. Drone-based surveillance

Conventional techniques of VSDM rely on fixed surveillance using monocular and intelligent cameras, which can only monitor a specific area. In contrast, drone-based surveillance is flexible, convenient, and broad in coverage. Drone-based VSDM analysis is a better option, as it can help monitor scenes from different points of view (Kadam, Seshapalli, Nayak, & Shaikh, 2021; Kumar et al., 2021) However, drones’ images show complex backgrounds because of varying scenarios, altitudes, diversity, and illumination. These complex backgrounds have significant interference with the VSDM. Typically, quickly and accurately detecting individuals is challenging in such conditions. More attention can be paid to the targets in complex backgrounds. Recent studies have proved that the spatial attention mechanism can achieve this goal. Specifically, it has been demonstrated that spatial attention enhances the features we are interested in and ignores unimportant characteristics. Using drones for VSDM and other monitoring applications (e.g., face mask detection) is getting increasing attention because of their flexibility, although the computing power and memory capabilities are limited during timely distance monitoring. In this regard, performing real-time drone-based VSDM is a major issue. The Landing AI Company (Social distancing detector, 2022) developed a real-time VSDM solution that (i) detects pedestrians in video streams recorded with drones and (ii) uses the BEV of frames to measure physical distances between individuals. For instance, Ramadass, Arunachalam, and Sagayasree (2020) use a drone for VSDM of face mask detection of people in a public place. If violations are detected, the drone sends alarms to the nearby police station and provides the public with alerts. It can also carry and drop face masks to individuals. Similarly, in Kadam et al. (2021), Shao et al. (2021), autonomous drones are used for VSDM (Kadam et al., 2021).

### 3. Datasets

To validate the VSDM algorithms in crowded areas, various publicly available video surveillance datasets have been used. Typically,
most of these datasets have already been employed to validate different video surveillance tasks, such as pedestrian detection, motion detection, crowd management, abnormal event detection, etc. For instance, Shorfuazzaman, Hossain, and Alhamid (2021) uses the Oxford town center (OTC) dataset (Benfold & Reid, 2011), which has been released by Oxford University. It encompasses one video sequence recorded in a semi-crowded urban street at a sampling rate of 25 frames per second (FPS) and with a resolution of 1920 × 1080. The ground truth BBs of the pedestrians are also generated in all the frames. In Su et al. (2021), in addition to realizing a new VSDM datasets, namely SCU-VSD, two other datasets are considered, i.e., Market1501 (Zheng et al., 2015) and MOTT16 (Milan, Leal-Taixé, Reid, Roth, & Schindler, 2016). Typically, SCU-VSD is a data repository including 8 video sequences that have been recorded from the pedestrian street. They have a sampling rate of 25 fps, a duration of 60 s, and a resolution of 1920 × 1080 with numerous scenes and perspective views. Market1501 includes images recorded in front of a supermarket (Tsinghua University) and encompasses 12,936 images for training and 3,368 images for test. MOT16 includes 7 videos employed for training and verification and another 7 for the test with a resolution of 1920 × 1080 and 640 × 480. It contains as well top-view scenes recorded with a surveillance camera and front-view scenes collected with a moving camera. The varying illumination, number of pedestrians, and complex scene have made this dataset very challenging for VSDM applications.

Shrestha et al. (2020) train their VSDM system on the PASCAL visual object classes challenge (VOC) 2007 (Everingham et al., 2008) and VOC 2012 (Everingham & Winn, 2011) datasets. Next, the system is tested on the PASCAL VOC 2007 test set. Specifically, VOC 2007 and VOC 2012 include 9963 and 11,540 images with objects over 20 different classes. In this case, the system performance has been reported only for the person class. Shao et al. (2021) validate their real-time drone-based VSDM system using a merge-head dataset. It includes 18,767 video frames recorded at a resolution of 1920 × 1080 and divided into a training set (15,940 frames), validation set (1340 frames), and test set (1487 frames).

In Al-Sa’d et al. (2022), EPFL-MPVT (Fleuret, Berclaz, Lengagne, & Fua, 2007), EPFL-Wildtrack (Chavdarova et al., 2018), and OTC (Benfold & Reid, 2011) datasets are used to evaluate DL models. The EPFL-MPV comprises four video sequences of six individuals freely moving in a room. Different scenes are collected from different points of view. Each video includes 2954 frames recorded at a sampling rate of 25 fps with a resolution of 920 × 1080. Besides, EPFL-Wildtrack comprises 7 video sequences of 400 frames each, describing the movement of 20 pedestrians outside the principal building of the ETH university (Switzerland). Pedestrians scenes have been collected using different cameras installed at different points of view. At the same time, OTC includes one video sequence collected on a pedestrian, which has 4501 frames recorded using a unique camera at 25 fps. In Madame et Chitre (2021), the INRIA person dataset (Dalal & Triggs, 2005) is utilized for training the DL models, which contains training and testing data and their corresponding annotations. While for inference or testing, the OTC dataset is considered along with the performance evaluation of tracking and surveillance (PETS 2009) dataset, in which both contain numerous crowd activities. Specifically, PETS contains video frames for different purposes, such as people tracking, crowd density estimation, flow analysis, etc. In Rahim, Magbool, and Rana (2021), Rahim et al. use the exclusively dark (ExDark) image dataset (Loh & Chan, 2019) for validating their VSDM solution. ExDark includes 12 different classes of objects with annotations, while only the pedestrian detection class has been considered for training developed algorithms. Moreover, it encompasses various indoor and outdoor low-light images. In Ni, Nath, Sampathkumar, and Behzadan (2021), to avoid overfitting, the YOLO-based VSDM solution is trained on multiple datasets, including PnP-FudanPed (Wang, Shi, Song, Shen, et al., 2007), which is a small dataset comprises one video showing 170 pedestrians walking on a street. Because it is not sufficient to train DL models, other large-scale datasets are utilized, e.g., VOC 2010 (Everingham, Van Gool, Williams, Winn, & Zisserman, 2010), and Microsoft common objects in context (MS-COCO) (Lin et al., 2014). Typically, the VOC 2010 dataset includes 20 different object classes; however, only the pedestrian class is considered to validate VSDM systems. More specifically, only that illustrate individuals riding bikes, running and/or walking are used.

In Yang, Yurtsever, Renganathan, Redmill, and Ozguner (2021), three pedestrian crowd dataset are employed to evaluate an SSD-based VSDM system, i.e., OTC, Mall Dataset (Mall-D) (Chen, Loy, Gong, & Xiang, 2012), and train station dataset (TSD) (Zhou, Wang, & Tang, 2012). The Mall-D is a 2000 video frames dataset with a resolution of 320 × 240, which is proposed originally for crowd counting. TSD is a one-video dataset comprising 50,010 frames, which is collected with a 25 fps rate and a resolution of 480 × 720. In Rezaei and Azarni (2020), Rezaei et al. validate their VSDM system using multi-object annotated datasets, i.e., VOC 2010 (Everingham et al., 2010), COCO, ImageNet (Russakovsky et al., 2015), and Google Open Images (GOI) datasets V6+ (Kuznetsova et al., 2020). The last one has 16 Million ground-truth BBs from 600 groups. Only the classes corresponding to human detection and identification are used. It is also a labeled dataset, where the BB labels have been used on every image and the corresponding coordinates of every label. In Shareef, Yannawar, Abdul-Qawy, and Ahmed (2022), in addition to Mall-D, PETS 2009, and OTC, Shareef et al. use the VIRAT (Oh et al., 2011), which is a natural, realistic, and challenging video surveillance dataset.

Overall, it is worth noting that most existing VSDM have been validated on already existing datasets that have been proposed for validating different video surveillance tasks, such as object detection, human action recognition, multi-object tracking, etc. This is mainly due to the similarities between those tasks and the VSDM task and the open challenges presented by these comprehensive and public datasets. On the other hand, very few datasets have been launched to particularly validate VSDM algorithms, such as SCU-VSD (Su et al., 2021).

4. Non-visual social distancing monitoring (NVSDM)

Recent progress in AI, the Internet of things (IoT), and wireless communication has enabled the collection of real-world data corresponding to human behavior and social contact. In this respect, many SDM tools have been proposed to track people in public areas using WiFi (Faggian, Urbani, & Zanotto, 2020) and Bluetooth (Berke et al., 2020) traces. Additionally, different smartphone apps, such as (Braithwaite, Callender, Bullock, & Aldridge, 2020; Cho, Ippolito, & Yu, 2020; Mbane, 2020; Udugama et al., 2020), active radio frequency identifications devices (RFID) or other wireless sensors (Bianco, Panunzio, & Marrocco, 2021; Fazio, Buzachis, Galletta, Celesti, & Villari, 2020; Nguyen et al., 2020; Zhang et al., 2020) have also been developed and used to motivate SD during the COVID-19 pandemic. Typically, these tools have been utilized to glean data on the proximity of human-to-human interactions.

For instance, in Bian, Zhou, Bello, and Lukowicz (2020), a wearable, oscillating magnetic field-based proximity sensing system is proposed for monitoring SD. It can track the individual’s SD in real-time and enjoy better reliability than Bluetooth RSSI signal-based SD tracking solutions. In Oransirikul and Takada (2020), SD warnings are generated based on separating passing individuals from waiting individuals. Precisely, the activity of Wi-Fi signals from mobile devices has been passively monitored to check that the number of individuals in a specific area has exceeded the allowable density. If yes, individuals are provided with warnings to keep SD. In Chandel, Banerjee, and Ghose (2020), a mobile-based platform for monitoring SD in enterprise scenarios named “ProxiTrak” is proposed. It aids in tracking the path of potential COVID-19 transmission among an ensemble of individuals. Additionally, it helps guide the individuals to follow SD rules by providing real-time alerts on their mobile phones once they violate SD norms or are exposed to a person who has tested positive. In this regard,
a classification algorithm is devised for making proximity decisions on the mobile phone itself using received signal strength indicator (RSSI) data of the on-board Bluetooth low energy (BLE) module. Besides, in Li, Sharma, Mishra, Batista, and Seneviratne (2021), Li et al. address the SD problem by developing a non-intrusive approach that monitors physical distances within a given space based on channel state information (CSI) from passive WiFi sensing. In this context, the frequency selective behavior of CSI is exploited by a support vector machine (SVM) classifier to improve the accuracy of SD detection and crowd counting.

Also, it is worth noting that many countries have used the global positioning system (GPS) to record the activities of people who tested positive. This helps track their traces and observe the probabilities of their contacts with fit people. For example, the EHTERAZ app has been used by the Qatar government to monitor and track doubted or diseased people and guarantee that they comply with the COVID-19 precautions (El-Haddadeh, Fadlalla, & Hindi, 2021). While in India, the Arogya setup app has been utilized by the government, which employs Bluetooth and GPS to localize and monitor COVID19 patients in public areas (Sharan, Chanu, Jena, Arunachalam, & Choudhary, 2020). However, most of these apps are only appropriate for indoor environments, and their accuracy significantly drops in dynamic environments. Moreover, they have significant privacy issues, and scalability problems (Borra, 2020).

Although NVSDM systems can help accurately detect physical distances between pedestrians, they are usually based on sensors that are handed out to achieve such a task. This can act as a medium for spreading the virus.

5. Visual social distancing monitoring (VSDM)

Since the outbreak of the COVID-19 pandemic, a large number of frameworks based on AI have been proposed to help fight against the virus. The literatures on VSDM around the world are arising. Various journal special issues and many international conferences were organized with many solutions introduced for resolving the VSDM problem in the last two years. This section sheds light on the state-of-the-art VSDM techniques. Typically, VSDM frameworks can be classified with reference to different aspects, such as the adopted model (conventional ML or DL), feature extraction (hand-crafted or neural network), data recording methodology (fixed or drone-based), complexity object detectors (complex or lightweight), object detection stages (single-stage or multi-stage), etc. Fig. 4 illustrates the proposed taxonomy.

5.1. Hand-crafted feature-based methods

In Cristani, Del Bue, Murino, Setti, and Vinciarelli (2020), a VSDM approach that relies on body pose estimation is introduced, where the
body pose detector has been utilized for detecting visible pedestrians. Then, after converting the video frames into a top view (BEV) representation, every detected person is considered the center of a circle, while the radius represents the safe distance. In this regard, the VSDM task problem has been transformed into a sphere collision problem. In Al-Sa’d et al. (2022), a VSDM and crowd management system is introduced, which is based on (i) detecting pedestrians using a global nearest neighbor tracking (GNN), which is a real-time light-weight MOT approach (based on allocating detection/prediction annotations to tracks, (ii) and preserving their track records), (ii) filtering region of interest (ROI), (iii) transforming video frames into a top-View, (iv) tracking and smoothing, (v) estimating parameters, and (vi) detecting (SD) violations. In Aghaee et al. (2021), a semi-automatic VSDM approach is proposed for approximating the homography matrices between the image plan and scene ground. Using the measured homography, an off-the-shelf pose detection is then leveraged for detecting body poses on images and reasoning upon their interpersonal distances using the length of their body parts. Moving on, interpersonal distances are examined to identify potential SD violations.

In Ziran and Dahnoun (2021), Ziran et al. propose a contactless and real-time solution to monitor SD using stereo cameras, where pedestrians are first detected using a histogram of gradients (HOG) in the reduced ROIs of each frame. Moving on, a disparity map is generated for regions of the image with detected people before calculating the distances between detected persons using the hypotenuse theorem. In Jayatilaka et al. (2022), an end-to-end VSDM method is developed based on graph theory. Typically, a temporal graph representation structurally stores the information extracted by the object detector. Specifically, individuals are represented by nodes with time-varying properties for their location and behavior. The edges between people represent the interactions and social groups. Next, the graphs are interpreted, and the threat levels in each are quantified based on primary and secondary threat parameters, including proximity and group dynamics extracted from the graph representation and individuals’ behavior.

5.2. CNN-based VSDM

CNNs have recently been considered a major player in different research topics, such as feature extraction, object detection, image segmentation, and human detection. Moreover, developing extended memory capacities and faster CPUs and GPUs have enabled the computer vision community to create powerful and robust pedestrian detectors that significantly outperform traditional ML algorithms. Despite that, many challenges still persist, including detection accuracy, detection speed, and computational training cost. These challenges also apply to the VSDM problem and need to be resolved to develop efficient and real-time VSDM systems. For VSDM, the pedestrian detection stage is the most critical part, and distance measurement accuracy depends mainly on it. To that end, most of the contributions have been focused on developing accurate people detectors using CNNs. The latter can be divided into single-stage CNN-based detectors, two-stage CNN-based, and multi-object detectors.

5.2.1. Single-stage CNN-based pedestrian detectors

A one-stage CNN-based detector relies on a unique pass through the CNN model to predict all the BBs in one go. This is appropriate for implementing mobile devices, such as drones, as it is fast. The most famous examples of one-stage CNN-based detectors are SSD, YOLO, RetinaNet, DetectNet and SqueezeDet (Faragallah et al., 2022).

YOLOv1: it is based on framing the pedestrian detection as a regression problem, and hence spatially separating BBs and associated class probabilities. Few frameworks have been developed using this architecture. For YOLOv1 instance, in Mercaldo, Martinelli, and Santone (2021), a YOLOv1 object detector is employed to detect people before using Euclidean distance between the people centroid to quantify the distance. Similarly, in Anitha Kumari, Purusoithaman, Dharani, and Padmashani (2021), VSDM approach based on YOLOv1 is developed and implemented on a Jetson Nano computing board. In Mercaldo et al. (2021), the YOLOv1 object detector is employed to detect people before using Euclidean distance metric to quantify the physical distances between people’s centroids.

YOLOv2: it is built upon the DarkNet-19, which is the model backbone. Compared to YOLOv1, YOLOv2 relies on removing fully connected layers and using anchor boxes for predicting BBs. Saponara, Elhanashi, and Gagliardi (2021) develop a VSDM scheme based on YOLOv2, which is applied to video streaming from thermal cameras. This approach enables tracking people, detecting SD violations, and monitoring body temperature. Moreover, the developed solution has been implemented on a Jetson Nano, which includes a fixed camera before testing it in a distributed surveillance system for visualizing individuals from multiple cameras in a centralized manner.

YOLOv3: it utilizes the complex DarkNet-53 as the model architecture. In Ramadass et al. (2020) an autonomous drone-based VSDM is proposed, in which the YOLOv3-based pedestrian detector has been trained on a dataset includes images of side and frontal views for a large number of people. This study has also been extended to detect face masks. The developed algorithm was then implemented on a surveillance drone with a camera to detect the physical distances between pedestrians from the frontal and side views. Similarly, the authors in Sathyamoorthy, Patel, Savle, Paul, and Manocha (2020) develop a pedestrian detection approach for VSDM in crowded areas using the YOLOv3-based detector designed in Wojke et al. (2017). Typically, a robot augmented with an RGB depth camera with 2D lidar is monitored in crowd gatherings for performing collision-free navigation. Moving on, YOLOv3 is utilized in Yang, Yurtsever, et al. (2021) to detect pedestrians in video sequences and identify SD violations. Specifically, the BEVs have been adopted to estimate the distances between pedestrians. Additionally, the density of crowd gatherings has been estimated to alert for critically dense areas.

In Magoo, Singh, Jindal, Hooda, and Rana (2021), a BEV VSDM scheme based on YOLOv3 object detection model is introduced. Typically, key feature patterns are detected using a key point regressor. Moreover, once a massive crowd is detected, BBs are used for detecting the individuals violating the SD norms. In Ahmed, Ahmad, Rodrigues, Jeon, and Din (2021), an SD tracking system is developed based on YOLOv3 object recognition paradigm, which helps (i) detect humans in video streams, and (ii) measure SD violations between people by approximating physical distances to pixels and setting an empirical threshold. Additionally, a transfer learning scheme is utilized to overcome the problem of data scarcity and improve the model’s accuracy. Using the same approach, the authors in Shalini, Margret, Niraimathi, and Subashree (2021), Widiatmoko, Berchmans, and Setiawan (2021) calibrate videos in the BEV plan before feeding them as inputs to the pre-trained YOLOv3 model. However, both studies do not provide enough assessment results.

In Pi et al. (2021), the study focuses on contact tracing using CNN to generate quantifiable metrics. Typically, a YOLOv3 network has been run on a training labeled video dataset, including pedestrians. Afterward, the trained architecture is validated on real-world crosswalk video sequences collected during the start of the pandemic in Xiamen, China. Following, identified pedestrians are projected onto an orthogonal map to trace contacts by (i) tracking movement trajectories and (ii) simulating the spread of droplets among the healthy population. Non-maximum suppression and Network pruning have been used to optimize model performance, resulting in an average precision of 69.41%.

In Hou, Baharuddin, Yussof, and Dzulkifli (2020), the pre-trained YOLOv3 is utilized for pedestrian detection in video sequences. Then, video frames are transformed into a top-down view to measure the
distances between individuals from the 2D plane. Moving on, non-compliant pairs of people are identified with a red frame and red line. Besides, Shorluzzaman et al. (2023) introduce a YOLOv3-based VSDM system and compare its performance with SSD and Faster-RCNN-based solutions. Accordingly, YOLOv3 has demonstrated its superiority in terms of the mAP score and speed (FPS).

YOLOv4: in Rodriguez, Luque, La Rosa, Esenarro, and Pandey (2020), a DL-crowd counting solution is developed for capacity control in commercial establishments buildings during the COVID-19 pandemic. It is based on YOLOv4 and has been validated on the MS-COCO dataset. Moreover, it can determine whether a person leaves or enters using the route and direction information, (ii) count remaining people inside a commercial building, and (iii) detect violations by comparing the result with a pre-defined threshold. However, the main drawback of this study is the lack of significant assessment. In Rahim et al. (2021), Rahim et al. propose a DL-based VSDM detection scheme based on the object detection YOLOv4 model. A fixed single motionless time of flight (ToF) camera is used to record video data. After people detection, the Euclidean distance metric is used to measure the physical distances between detected B Bs and then map them to real-world unit distance. Empirical evaluation has shown an mAP score of 7.84%, and the mean absolute error (MAE) between actual and measured social distance values has reached 1.01 cm. Similarly, in Ismail, Najeeb, Anzar, Aditya, and Poorna (2022), YOLOv4-based VSDM is proposed to detect pedestrians and then measure the distance between them using the Euclidean distance. This is to guarantee that people are properly following the SD norms.

In Ghasemi, Kostic, Ghaderi, and Zussman (2021), an accurate VSDM pipeline for automating video-based SD analysis, namely AutoSDA, is designed, in which the performance is insensitive to scene dynamics and the camera’s viewpoint. This method uses (i) a YOLOv4-based object detector and (ii) a people tracking approach based on Nvidia DCF-based tracker (NvDCF) for extracting pedestrian trajectories. The latter is then deployed for computing the proximity duration of every two unaffiliated pedestrians separately. In Shareef et al. (2022), a YOLOv4-based VSDM solution is developed that first detects pedestrians in video scenes before a predefined SD threshold and a violation index to detect SD violations. Moving on, warnings are produced to make immediate awareness actions. Because low-light environments can result in spreading COVID-19, developing efficient VSDM schemes that address this issue is of utmost importance. To that end, Rahim, Maqbool, Mirza, Afzal, and Asghar (2022) introduce DeptTSol, which is a CSP-ized YOLOv4-based VSDM system under different light conditions. It also enables the monitoring of pedestrians at varying camera distances.

SSD: in Khel et al. (2021), the authors employ a lightweight CNN-based Mobilenetv2 architecture as a framework for the classifier to detect face masks and monitor SD. Additionally, an SSD is used to extract relevant features, while a spatial pyramid pooling (SPP) is deployed for integrating the collected features and improving the model’s accuracy. Similarly, in Qin and Xu (2021), SSD300-based VSDM that is built upon a feed-forward convolutional network (FFCN) is proposed. It produces a fixed-size collection of BBs and scores for the presence of pedestrians, then estimates the distance between them using the Euclidean function. In Gopal and Ganesan (2022), an SSD-based VSDM scheme is introduced, where an overhead position dataset and a pre-trained MS-COCO dataset have been used to train the pedestrian detector. Additionally, a transfer learning scheme has been employed to enhance the performance of the pre-trained model. A new layer has been integrated over the existing architecture to train the overhead dataset. Moving on, a centroid chasing algorithm, working on the concept of the fixed distance threshold, is deployed to identify people violating the SD norms.

RetinaNet: In Chaudhary (2020), a VSDM approach is developed and installed in the hardware of CCTV cameras for contact tracing. RetinaNet has been deployed to detect and track pedestrians before using the law of similar triangles to calculate the distance between them. Accordingly, 30% accuracy improvement has been achieved when the law of cosines is considered. Additionally, a multi-task cascaded CNN-based face detection has been utilized to identify people violating the SD norms. Besides, in Zhao et al. (2021), a VSDM approach is proposed based on obtaining pedestrian density and distance between each pedestrian pair. It uses three pre-trained CNN-based object detection architectures, i.e., RetinaNet, YOLOv3, and Mask R-CNN, as backbone models. Mask R-CNN and RetinaNet utilize ResNet101 as the network architecture, and these models are pre-trained using the MS-COCO dataset (Lin et al., 2014). Real-time video sequences gathered in New York City (NYC) have been employed to validate this framework. However, the performance has been quantified using the average pedestrian density (APD) and SD adherence rate (SADR), which cannot reflect the efficiency of the VSDM system.

To highlight the performance of one-stage pedestrian detectors for VSDM under different light conditions, seven detectors are evaluated on the ExDARK dataset to assess the accuracy and speed of their models, as explained in Rahim et al. (2022). Fig. 5 presents (a) the mAP performance at (i) various IoU thresholds (mAP_{IoU=0.5} and mAP_{IoU=0.75}) and (ii) considering different object dimensions (mAP_{small}, mAP_{medium} and mAP_{large}) and (b) the mAR performance with reference to (i) the detection number per image number (i.e. mAR_{max}=1, mAR_{max}=10) and mAR_{max=100}, (ii) the scale variation (mAR_{small}, AR_{medium} and AR_{large}). The CSP-ized YOLOv4 has outperformed all the other detectors for both mAP and mAR. The CSP-ized YOLOv4 has achieved the best
performance in terms of both the mAP and mAR scores compared to the six other one-stage detectors. For instance, up to 99.7% mAP has been reached by CSP-ized YOLOv4 under mAP$_{IOU=0.5}$. For the computational cost, the test performance of the processed frames per second of each model has been assessed on a Tesla T4 GPU, which has a 512 × 512 network size, as portrayed in Fig. 6. It has been shown that the best performance has been reached by the CSP-ized YOLOv4, where 51.2 fps has been attained. Overall, one-stage pedestrian detectors have received increasing attention for VSDM due to their computational efficiency and competitive detection performance.

5.2.2. Two-stage CNN methods

RCNN: in Degadwala et al. (2020), different DL architectures are used to address the VSDM problem, including RCNN, Faster-RCNN, SSD, YOLOv1, YOLOv2, and YOLOv3. After detecting people in the video frames from MS-COCO (Lin et al., 2014) and PASCAL-VOC (Everingham & Winn, 2011) datasets, Euclidean distance has been considered to quantify the distance between them.

Faster-RCNN: it improves some of the problems of RCNN and provides a faster architecture for pedestrian detection. In Saponara et al. (2021), Faster-RCNN has been implemented to perform a VSDM. Its performance has been compared with YOLOv2 and YOLOv4-tiny. The latter one has the best performance in terms of pedestrian detection accuracy and computation efficiency.

Faster-RCNN: it is built on RCNN and Fast-RCNN by using a region proposal network (RPN) for sharing complete images’ convolutional characteristics with a detection network, which helps generate almost cost-free region proposals. In Ahmed, Ahmad, and Jeon (2021), a transfer-learning-based Faster-RCNN is introduced to detect persons in video frames using BBs, which have been recorded in top view environments. Typically, a pre-trained model has been combined with a new trained layer. Moving on, Euclidean distance is considered to estimate the distances between detected individuals. After estimating the central point of a BB, a distance to pixel threshold is set to determine whether individuals respect SD or not.

In Sabraoui et al. (2020), a DL-based VSDM based on the social internet of vehicles (StIoV) named DeepDist is proposed to detect SD violations in real-time. Typically, the Faster-RCNN model is utilized for detecting physical distancing violations between objects in video sequences recorded with vehicles equipped with thermal and vision imaging systems. The performance of this approach is evaluated on the Stanford vehicles’ dataset (SVD), network simulator (NS-3), and the simulation of urban mobility (SUMO). Similarly, in Shah, Chandaliya, Bhuta, and Kanani (2021), a pre-trained Faster-RCNN is selected to perform VSDM from videos recorded using CCTV Cameras. In Tanwar et al. (2021), the VSDM task is performed using Faster-RCNN and YOLOv2 to analyze videos recorded using drone-based and CCTV cameras. The Euclidean distance has been utilized to calculate the distance between pedestrians. More importantly, the developed VSDM solution is augmented with a privacy preservation module based on blockchain, which helps ensure trusted and secure data exchange between different entities and the surveillance center at the physical layer. Additionally, blockchain currencies are utilized to pay fines if individuals violate SD norms.

Mask-RCNN: this architecture helps extend and improve Faster-RCNN (i) using the ROI align instead of an ROI pooling to address the location misalignment problem existing in the ROI pooling, and (ii) through the addition of a mask branch. However, few VSDM frameworks have been designed based on Mask-RCNN. For instance, Gupta, Kapil, Kanahasabai, Joshi, and Joshi (2020) develop a Mask-RCNN-based VSDM by (i) detecting pedestrians in each video frame, (ii) splitting the input proposals from the region proposal network (RPN) into “bins” using bi-linear interpolation, and (iii) applying a pairwise distance measurement to detect if the SD requirements respected.

EfficientDet: in Madane and Chitre (2021), three pre-trained object detectors, namely EfficientDet-D0, EfficientDet-D5, and DETR, having ResNet-50 as a backbone, are used to detect pedestrians in public areas. Moving on, the fine-tuned models have been evaluated on OTC (Davis & Sharma, 2007) and PETS (Ferryman & Shahrokhi, 2009) people tracking datasets. In this respect, the developed VSDM system has been built upon DETection Transformer (DETR) with the aid of a perspective transform and camera calibration. This makes the distancing monitoring approach independent of the camera angle or position.

Other models: it is worth mentioning that there are other frameworks that have used other object detectors, such as Ghodgaonkar et al. (2020), where Cascade-HRNet is deployed to detect pedestrians after being trained on the crowd human dataset (Shao et al., 2018). In Dai et al. (2021), Dai et al introduce BEV-Net, a multi-branch network that localizes pedestrians in real coordinates and identifies high-risk areas of SD violation. Typically, this network aggregates camera pose estimation, feet, and head location detection, a differentiable homography scheme for mapping images into BEV coordinates, and uses geometric reasoning for producing BEV maps of individuals’ locations in the s

5.2.3. Lightweight CNN models

In contrast to most of the studies that have focused on a front or side perspective for social distance tracking, a BEV is adopted to track SD in Karaman et al. (2021), where a lightweight CNN model, i.e., MobileNet with SSDv3 and other complex CNN models, i.e., Faster-RCNN (with ResNet-50), Faster-RCNN (with Inception-v2), are deployed to detect people in video sequences. A prototype has also been developed by implementing the Faster-RCNN-based image analysis algorithm on an embedded Jetson Nano platform, including a Raspberry Pi camera. Moreover, the system has been tested in various public spaces, where audible and light warnings have been used to detect social distance violations. Another VSDM scheme is introduced in Khandelwal et al. (2020) using MobileNetv2 network as a lightweight person detector to alleviate the computational cost, showing less accuracy in comparison with other common models. The Euclidean distance between detected people has been measured using a symmetric distance matrix and a 3D projected image of each frame. Moreover, this approach only focuses on an indoor manufactory-setup distance measurement and does not provide any statistical assessment on the virus spread. In Ansari et al. (2021), a VSDM using a compact CNN-based sequential model is proposed to first detect pedestrians in video frames collected using CCTV cameras. In doing so, a sliding window concept has been adopted as a region proposal when detecting pedestrians in each frame. Next, Euclidean distance has been used to measure the physical distance between detected persons.
SocialdistancingNet-19 includes two subnetworks used for feature extraction and detection: CNN and MobileNet-V2 models. Moreover, performance has been compared to reduced ResNet-50, and ResNet-18 architectures, where an accuracy of 92.8% has been reached by SocialdistancingNet-19. In Shao et al. (2021) the lightweight PeleeNet model is used as a backbone for a pedestrian detection module implemented on drones. This enables detecting pedestrians in real-time based on human head detection on UAV images. Typically, spatial attention and multi-scale features are easily incorporated to enhance small objects’ features, such as human heads. After that, SD is measured based on human head detection on UAV images. Typically, spatial attention and multi-scale features are easily incorporated to enhance small objects’ features, such as human heads. After that, SD is measured based on human head detection on UAV images. Typically, spatial attention and multi-scale features are easily incorporated to enhance small objects’ features, such as human heads. After that, SD is measured based on human head detection on UAV images. Typically, spatial attention and multi-scale features are easily incorporated to enhance small objects’ features, such as human heads. After that, SD is measured based on human head detection on UAV images. Typically, spatial attention and multi-scale features are easily incorporated to enhance small objects’ features, such as human heads. After that, SD is measured based on human head detection on UAV images. Typically, spatial attention and multi-scale features are easily incorporated to enhance small objects’ features, such as human heads. After that, SD is measured based on human head detection on UAV images. Typically, spatial attention and multi-scale features are easily incorporated to enhance small objects’ features, such as human heads. After that, SD is measured based on human head detection on UAV images. Typically, spatial attention and multi-scale features are easily incorporated to enhance small objects’ features, such as human heads. After that, SD is measured based on human head detection on UAV images.
### 5.2.4. Multi-object tracking (MOT)

Besides, IMPERSONAL is introduced in Giuliani et al. (2021) to detect and track SD and alert users in case of gatherings. The process is conducted in three steps: i) object detection, multi-object tracking (MOT), and (iii) distance estimation. This system is built upon Fair-MOT (Zhang, Wang, Wang, Zeng, & Liu, 2021), which is an MOT scheme, which is in turn based on a ResNet-34 backbone network. Moving forward, the retrieved information is then sent to an IoT subsystem to (i) identify the anonymous IDs of people belonging to a gathering and (ii) provide them with alert messages. This framework has been validated on PET52006 datasets, (iv) work on fixed or moving cameras.

#### Table 1 (continued).

| Work                          | ML model  | Description                                                                 | Dataset                                      | Best VSDM performance | Advantage/limitation                                                                 |
|-------------------------------|-----------|-----------------------------------------------------------------------------|----------------------------------------------|------------------------|-------------------------------------------------------------------------------------|
| Shareef et al. (2022)         | YOLOv4    | A predefined SD threshold and a violation index are used to detect SD violations determine when the violation | Mall-D, PETS2009, OTC, VIRAT                | Acc = 96%              | Privacy preservation was not considered and cannot distinguish between individuals from the same family and strangers. |
| Anvari, Singh, et al. (2021)  | CNN       | Real-time VSDM                                                             | INRIA                                        | Acc = 98.50%           | Pedestrians overlapping can bias detection performance (unique camera) and validation on a small image dataset. |
| Saponara et al. (2021)        | YOLOv2    | Real-time VSDM and body temperature detection from thermal videos.          | Private thermal image dataset                | Accuracy = 95.6%       | (i) Real-time validation in real-world scenarios, (ii) appropriate for distributed video surveillance system. |
| Magoo et al. (2021)           | YOLOv3    | DL-based BEV SD analysis                                                   | OxTown                                       | mAP = 93.6%            | Extremely sensitive to the spatial position of the camera.                            |
| Valencia et al. (2021)        | Tiny-YOLOv4 and DeepSORT | Crowd counting and SD monitoring in a top-view camera perspective | Youtube video                              | mAP = 92.94%           | Process video streaming in real-time, however, further improvement is needed to improve detection accuracy. |
| Giuliani, Innocenti, Mazzenga, Vegni, and Vizzarri (2021) | ResNet-34 | Use CV and radio IoT sub-systems for tracking people and retrieving the IDs of their devices. | PETS2006 and real-world video data          | Acc = 95.2%, F1 = 97.5% | Raise some privacy issues.                                                              |
| Bertoni, Kreiss, and Alahi (2021) | DFCN    | A cost-effective VSD approach that perceives people's 3D locations and their body orientation from images | KITTI dataset                               | Acc = 84.7%, recall = 85.3% | Work with single RGB images, (ii) privacy safe, (iii) does not require homography calibration, (iii) generalize well across different datasets, (iv) work on fixed or moving cameras. |
| Rahim et al. (2021)           | YOLOv4    | Validation on video data recorded using fixed single motionless time of flight (ToF) camera | ExDARK dataset                              | mAP = 97.84%, MAE = 1.01 cm | Can be applied in real-world scenarios because of high precision and the low error rate. Used only with fixed cameras. |
| Sahraoui et al. (2020)        | Faster-RCNN | Using SloV to detect SD violations in real-time                            | Stanford Vehicles' Dataset                  | mAP = 0.76             | Validation on open-source simulation platforms, and further improvement are required to improve the detection accuracy. |
| Pi et al. (2021)              | YOLOv3    | Contact tracing and simulation the spread of droplets among the healthy population. | PennFudanPeel, MS-COCO, VOC                  | Precision = 69.41%     | Do not report SD violations. Average precision is quite low. |
| Hou et al. (2020)             | YOLOv3    | Transforming video frames into top-down view for distance measurement.     | Private data                                 | N/A                    | Validation on small dataset and the performance was not reported.                    |
| Ghasemi, Kostic, et al. (2021) | YOLOv4    | People tracking using NvDDCF to extract pedestrian trajectories.           | SVD, NS-3, SUMO                              | N/A                    | No insight was provided about the accuracy of SD detection.                           |
| Shao et al. (2021)            | PeleeNet  | Real-time UAV-based VSDM using light-weight CNN                            | Merge-Head, UAV-Head                         | AP = 92.22%            | Instability caused intensive wind and the performance needs further improvement.       |
| Tanwar et al. (2021)          | Faster-RCNN, YOLOv2 | Secure and privacy preserving VSDM using blockchain                  | COCO                                         | AUC = 73%              | Although a secure and privacy preserving VSDM framework is presented, the detection accuracy needs further improvement. |
tracking (SORT) tracking technique. This has resulted in efficient pedestrian detection and SD analysis. The overall system has been trained on MS-COCO and GOI datasets and validated on the OTC dataset and real-world scenarios with challenging conditions, e.g., different lightning rates, occlusion, and partial visibility. Concretely, a 99.8% mAP and 24.1 fps processing have been achieved. Moving on, statistical analysis has been used to assess online infection risks using SD violations and spatio-temporal information from pedestrian movement trajectories. Fig. 7 presents the flowchart of the YOLOv4-based VSDM proposed in Rezaei and Azarmi (2020).

5.3. Transfer-learning-based VSDM

TL consists of training a model on a specific domain (or task) and then transferring the acquired knowledge to a new, similar environment (or task). For example, let us consider pedestrian detection, where a DL algorithm can be pre-trained on the large-scale ImageNet dataset to generate optimal model parameters. Next, a part of the model is re-trained (i.e., fine-tuning), and the validation process is performed on a new video target dataset collected from a real-world scenario (Ahmed, Jeon, Chehri, & Hassan, 2021; Loey, Manogaran, Taha, & Khalifa, 2021). Additionally, DL models can be pre-trained to perform a specific task like generic object detection on large-scale datasets, such as ImageNet, and fine-tuned to conduct a different but related task, such as pedestrian detection in VSDM. Fig. 8 explains the difference between conventional ML and TL techniques.

5.3.1. Fine-tuning

Most VSDM-based TL techniques are based on fine-tuning a pre-trained DL model when the source and target domains are almost similar. In Shin and Moon (2021), Shin et al. firstly detect pedestrians using a YLOLOv4-based TL object detector in CCTV images. After that, DeepSORT-based MOT is utilized for assigning IDs and tracking objects. Moving forward, the weights of the transformation matrix are derived to extract the object coordinates using image warping of the initial frames. The center points of the BBs for the pedestrians are transformed to fit the shapes of the transformed frames using the extracted transform matrix weights. Following, actual distances are calculated using the Euclidean distance function. Punn et al. (2020) combine fine-tuned YOLOv3-based VSDM approach for detecting pedestrians, and Deepsort technique (Wojke et al., 2017) that aims at tracking detected persons using assigned IDs and BBs. To fine-tune the pedestrian detector, an open image dataset (OID) has been considered while the validation has been conducted on the OTC dataset. Moving forward, the empirical results have been compared with SSD and Faster-RCNN. However, no discussions about the validity of SD measurements are provided, and the statistical analysis of the obtained results is missing.

Using the same process in Ahmed, Ahmad, Rodrigues, et al. (2021), SD tracking is performed by detecting people in video sequences using a YOLOv3 object recognition system. Also, a TL scheme is considered to reduce the computational cost and improve detection accuracy. Fig. 9 illustrates the flowchart of the TL-based pedestrian detection system using overhead video frames, which has been employed to measure the physical distances between pedestrians. Typically, fine-tuning is adopted by freezing all the layers of the pre-trained YOLOv3 architecture, and only one new layer is trained on the real-world video training set.

In Ahmed, Ahmad, and Jeon (2021), a transfer learning-based Faster-RCNN is introduced to detect persons in video frames using BBs, which have been recorded in top view environments. Typically, a pre-trained model has been combined with a new trained layer. Moving on, Euclidean distance is considered to estimate the distance between detected individuals. After catching the central point of a BB, a distance to pixel threshold is set to determine whether individuals
Fig. 9. The TL-based pedestrian detection framework proposed in Ahmed, Ahmad, Rodrigues, et al. (2021), which is built using YOLOv3 and overhead video frames from real-world.

respect the SD norms or not. In Bouhlel, Mliki, and Hammami (2021), a VSDM scheme using drone-based surveillance is proposed, which relies on crowd behavior analysis. Typically, crowd density is first estimated by categorizing the drone video frame patches into four classes: none, medium, sparse and dense. Next, pedestrians are detected and tracked before calculating their physical distances. A TL approach is adopted for crowd density estimation, where the pre-trained AlexNet is utilized. Typically, a fine-tuning is adopted by substituting the classification layer with a novel softmax layer to classify the crowd patches into the classes mentioned above. Three datasets have been used to validate this approach, including Mayenberg's dataset (Meynberg & Kuschik, 2013), Mliki's dataset (Hazar, Arous, & Hammami, 2019) and UCF-ARG (Agendran, Harper, & Shah, 2021).

5.3.2. Domain adaptation (DA)

DA refers to the possibility of applying a DL algorithm trained on a specific domain (source domain) to another distinct but related domain (target domain). This research topic has received increasing interest in the last decade as it helps in reducing the complexity of DL-based computer vision solutions (Khan & Alamin, 2021). Although the importance of DA, few VSDM frameworks have been built on it. For instance, the authors (Di Benedetto et al., 2022) propose a VSDM scheme to monitor compliance with SD norms in indoor and outdoor environments. In doing so, the DA-based VSDM strategy consists of (i) launching a new real-world crowd counting and monitoring dataset, namely CrowdVisorPisa; (ii) training a Faster-RCNN model on a synthetic dataset, namely Virtual Pedestrian Dataset (ViPeD) (Amato, Ciampi, Falchi, Gennaro, & Messina, 2019), to detect pedestrians; (iii) fine-tuning this model on real-world data by employing the balanced gradient contribution (BGC) method that helps mix synthetic and real-word data during the training to boost the performance; and (iv) measuring the physical distances between detected pedestrians using a pre-calibration strategy and a geometrical transformation.

Table 2 presents a summary of TL-based VSDM frameworks and their features concerning the adopted ML model, method description, datasets used for validation, best performance, and advantage/limitation. It has been seen that the best performance has been achieved by Bouhlel et al. (2021), where a TL-based AlexNet approach is adopted to perform VSDM in drone images. Typically, an accuracy of 99.58% has been reached.

5.4. 3D-based VSDM

The COVID-19 pandemic has shown the need to perceive people in 3D more than ever when using visual intelligence systems. In this context, efficiently monitoring SD requires not only going beyond a measure of distance but also perceiving people’s orientations and relative positions. Put differently, people talking to each other strongly influence the risk of contagion more than walking apart. To that end, Bertoni et al. (2021) develop a VSDM solution that analyzes SD based on both 3D localization and social cues. Typically, a DL-based VSDM method is proposed to detect people’s 3D locations and their body orientations from monocular cameras. Typically, this approach is built upon an improved version of MonoLoco (Bertoni, Kreiss, & Alahi, 2019), based on a deep fully-connected network (DFCN). Similarly, in Niu et al. (2021), Niu et al. introduce a 3D-based VSDM that enables detecting and localizing pedestrians in 3D using a combination of terrestrial point clouds and monocular images. Moreover, the correspondence between 2D image points and 3D world points has been used to calibrate the camera. Typically, point clouds have been utilized to extract the vertical coordinates of the ground plane (where the pedestrians stand). Moving on, the 3D coordinates of the pedestrian’s head and feet have then been estimated iteratively using collinear equations, assuming that the pedestrians are perpendicular to the ground. Therefore, this helps localize and determine pedestrians in 3D based on data from monocular cameras, which are broadly installed in smart cities.
5.5. Detection of free-standing conversation groups (FCGs) and social groups (SGs)

Seeking to prevent forming free-standing conversation groups (FCGs) social gathering, a convolutional variational autoencoder (CVAE) model is employed in Varghese and Thampi (2021) to develop a VSDM by integrating data from various sensor modalities. SD violations are detected, considering the spatial characteristics required for managing illumination variations and occlusions of video data. If SGs are detected as graphs using the pre-trained CVAE and connected components in graph theory, violation alerts are generated. Moreover, an SG graph clustering is performed using a cost function to identify FCGs based on a socio- psychological theory of Friends-formation. On the other hand, blind and visually impaired (BVI) people have some issues when practicing SD because of their low vision, which impedes them from maintaining a safe physical distance from other persons. To that end, the authors in Shrestha et al. (2020) introduced a smartphone-based VSDM based on CNN crowd detection before associating risks to BVI users via directive audio alerts on their mobile phones. Typically, pedestrians are first detected, and their distances from the mobile phone’s monocular camera feed are estimated. Moving on, pedestrians are clustered into crowds to calculate distance and density maps from the crowd centers. Lastly, the system tracks each detection in previous frames to create motion maps that help (i) predict the crowds' motion information and (ii) produce corresponding audio alerts. Active Crowd Analysis is designed for real-time smartphone use, utilizing the phone’s native hardware to ensure the BVI can safely maintain SD (Shrestha et al., 2020).

Moving on, in Usman et al. (2020), Usman et al. develop a VSDM for shopping malls using a crowd-based simulator. It is based on clustering consumers’ behavior into three levels and using agent control. The SD index (SDI) is introduced as an evaluation metric, which is estimated to indicate the tendency of consumers to maintain a safe distance during their shopping experience. Concretely, SDI represents the occupancy throughput and the number of detected SD violations. This simulated VSDM has been tested on different scenarios by varying navigational guidelines, occupancy rate, and agent behavior.

It is worth noting that apart from the aforementioned studies, various VSDM solutions have also been proposed to analyze SD between pedestrians during the pandemic. For instance, the ones developed by Trident (Face mask detection system using artificial intelligence, 2022) and Landing AI (Social distancing detector, 2022) use AI-based algorithms to measure the physical distances between pedestrians using surveillance cameras. In addition, some solutions utilize visual data recorded from LiDAR cameras (Social Distance Monitoring, 2022), and 3D cameras (Using 3d cameras to monitor social distancing, 2022) to control SD. Moreover, visual intelligence is also used for real-time face mask detection in public, such as Datakalab (Datakalab | Analyse de l'image par ordinateur, 2022), Trident (Face mask detection system using artificial intelligence, 2022) and Deloitte (Protected–your ai-solution for face mask detection in public places, 2022). These solutions provide an instant output, helping organizations meet public health guidelines.

6. Discussion and important findings

6.1. Pedestrian localization error

When developing VSDM systems, it is essential to assess the pedestrian localization errors that can occur due to different reasons, e.g., occlusions (as found in the Mall dataset), small sizes of pedestrians (as seen in TSD), noise, etc. However, most existing VSDM frameworks claimed that they had achieved a limited number of missed detections, which has slightly affected the monitoring of SD violations, as explained in Yang, Yurtseven, et al. (2021).

6.1.1. Indoor environment

A typical example of VSDM systems has been proposed in Niu et al. (2021), which enables the localization and detection of pedestrians in video frames recorded using monocular cameras before measuring the physical distances between them. Fig. 10 illustrates an example of an indoor scene at the CUMTB-Campus, where four pedestrians have been detected using YOLOv1. The pedestrians’ localization and height errors with different distances from the camera in this scene and the SD errors between adjacent pedestrians are evaluated. The results are reported in Table 3. Overall, it has clearly been seen that the most significant localization error has reached 0.32 m (pedestrian 1) while the most critical height error has attained 0.229 m (pedestrian 3). However, these errors have a slight effect on the SD monitoring errors, where the most significant error has attained 0.082 m (Pedestrians 3~4). Typically, if an SD norm of two meters is adopted, an average SD monitoring accuracy of 99.1% is reached.
Fig. 10. Example of an indoor video scene with four detected pedestrians recorded at the CUMTB-Campus to evaluate the VSDM system developed in Niu et al. (2021).

Table 3
Evaluation of pedestrian localization and height errors of an indoor video scene recorded at the CUMTB-Campus (Niu et al., 2021).

| Pedestrian number | Distance (m) | Localization error (m) | Height error (m) | Adjacent pedestrians | Truth value (m) | Measured value (m) | Absolute error (m) | SD accuracy |
|-------------------|--------------|------------------------|------------------|---------------------|-----------------|--------------------|-------------------|-------------|
| 1                 | 10.203       | 0.32                   | 0.054            | –                   | –               | –                  | –                 | –           |
| 2                 | 14.271       | 0.242                  | 0.007            | 1–2                 | 4.832           | 4.77               | 0.062             | 0.987       |
| 3                 | 24.421       | 0.133                  | 0.229            | 2–3                 | 10.266          | 10.199             | 0.067             | 0.993       |
| 4                 | 36.563       | 0.299                  | 0.071            | 3–4                 | 13.122          | 13.209             | 0.087             | 0.993       |
| Average           | –            | 0.248                  | 0.09             | –                   | –               | –                  | 0.072             | 0.991       |

Table 4
Evaluation of pedestrian localization and height errors in an outdoor scene recorded at CUMTB-Campus (Niu et al., 2021).

| Pedestrian number | Distance (m) | Localization error (m) | Height error (m) | Adjacent pedestrians | Truth value (m) | Measured value (m) | Absolute error (m) | SD accuracy |
|-------------------|--------------|------------------------|------------------|---------------------|-----------------|--------------------|-------------------|-------------|
| 1                 | 18.257       | 0.185                  | 0.058            | –                   | –               | –                  | –                 | –           |
| 2                 | 18.602       | 0.116                  | 0.043            | 1–2                 | 0.682           | 0.589              | 0.093             | 0.863       |
| 3                 | 25.197       | 0.476                  | 0.141            | 2–3                 | 8.059           | 8.176              | 0.117             | 0.985       |
| 4                 | 11.238       | 0.093                  | 0.017            | 3–4                 | 14.709          | 14.548             | 0.161             | 0.989       |
| 5                 | 51.072       | 0.517                  | 0.139            | 4–5                 | 41.132          | 40.925             | 0.207             | 0.994       |
| 6                 | 42.151       | 0.341                  | 0.114            | 5–6                 | 9.171           | 9.367              | 0.196             | 0.978       |
| 7                 | 41.984       | 0.392                  | 0.128            | 6–7                 | 6.076           | 6.558              | 0.482             | 0.918       |
| 8                 | 36.879       | 0.271                  | 0.122            | 7–8                 | 6.205           | 6.112              | 0.093             | 0.985       |
| Average           | –            | 0.298                  | 0.095            | –                   | –               | –                  | 0.140             | 0.945       |

6.1.2. Outdoor environment

The second example refers to assessing the pedestrian localization error in an outdoor scene recorded at the CUMTB campus (Niu et al., 2021). It includes eight pedestrians detected using YOLOv1, which are in different positions, including people overlapping, as portrayed in Fig. 11. Table 4 presents the localization and height errors of the pedestrians detected in this scene. Additionally, the SD monitoring errors between adjacent pedestrians are listed. It can be seen from the obtained results that the most significant localization error has been reached with pedestrian 5 since he is more than 50 m away from the camera. The second larger error has been obtained with pedestrian 3 (yellow box), mainly due to the occlusion issue. However, it is worth noting that the maximum relative error of SD monitoring is 0.207 m. Keeping in mind that an SD norm of two meters has been considered in this study, an average SD accuracy of 94.5% has been achieved.

Besides, in Shao et al. (2021), the error of pedestrian localization and accuracy of the VSDM system based on PeleeNet are evaluated under different scenes with a multitude of pedestrian position patterns. Fig. 12 portrays a typical scene (recorded with a drone-based camera) used to assess the VSDM system performance with eight pedestrian positions. Typically, detected social distances are compared with the ground truth before calculating each pedestrian pair’s absolute errors and SD accuracy. Table 5 depicts obtained results regarding the absolute errors (in m) and SD accuracy. Overall, an average error of 0.109 m has been achieved along with an SD accuracy of 0.945.

6.2. Critical discussion

The comprehensive overview conducted in this paper has shown that a significant amount of studies have been proposed to develop efficient VSDM systems and help slow down the spread of COVID-19.
Most of them are based on analyzing video sequences, detecting pedestrians in each frame, and quantifying the distances between detected people. From another point of view, most prototypes have focused on the side, and frontal camera perspectives, such as Ramadass et al. (2020) and Punn et al. (2020). Moreover, it has been demonstrated in the many frameworks that using visual data can effectively monitor SD by (i) accurately estimating physical distances, (ii) detecting crowd gatherings, and (iii) counting the number of people in each crowd. However, it is of utmost importance to mention that most existing VSDM solutions are based on a frame-by-frame SD analysis than on SD monitoring over time. In what follows, we summarize the main findings derived from this study.

**Datasets:** Validating VSDM frameworks necessitates at least one benchmark dataset, which includes a large number of images or video clips, different SD scenarios, different environments and scenes (outdoor, indoor (shopping malls, sports facilities, transport facilities, etc.)), and the proper ratio between virtual data and real data. However, datasets used for evaluating some existing frameworks have a set of limitations, which can be discussed as follows:

- Some datasets include a small number of images, e.g., several hundreds, which can limit the performance of DL algorithms trained on them and results in overfitting problems. Indeed, when DL models are used, a larger dataset often helps develop a more accurate DL model.
- Some datasets rely on simulated images/videos generated using virtual reality. DL algorithms are trained on these kinds of data, while in the real world, they should be validated on real images/videos. In this respect, their performance can be dropped due to the significant difference between the source and target domains.
- In some datasets, images/videos are gathered from simple scenes, which can easily bias them toward a special scene. In this regard, DL models trained on these datasets may be inefficient for new scenes.
- It has been demonstrated in the literature that training DL models on the VOC dataset for object detection (pedestrians) can improve models’ performance. Typically, while the mAP of a DL model can reach 35.9%–46.5% with the MS-COCO dataset, it can attain 57.9%–74.9% with the VOC dataset (Redmon & Farhadi, 2017). Therefore, this dataset has been used to pre-train different VSDM algorithms (Ahmed, Ahmad, Rodrigues, et al., 2021; Pi et al., 2021).
From another hand, although most developed systems are advantageous, they still have limitations and can be improved in different manners, including the (i) estimation of the bodies’ orientations for relaxing the assumption of vertically oriented subjects; (ii) fusion of pedestrian detection samples and distance measurements from multi-view cameras for assessing the environment state instead of the particular camera scenery; (3) development of online automatic training processes to track algorithms’ parameters; (4) integration of regression models for estimating crowd density maps; (5) detection of other abnormalities that can be related or not to the COVID-19 pandemic, e.g., smoke, fire or unattended objects in public areas, and any other abnormal events corresponding to crowd gatherings.

Additionally, even though some existing VSDM approaches have addressed both detection and tracking (e.g., Ahmed, Ahmad, Rodrigues, et al., 2021; Sathyamoorthy et al., 2020), the tracking schemes of these frameworks have been utilized to track detected pedestrians and then associate them with assigned IDs instead of using trajectory-based VSDM. Typically, these techniques based on frame-by-frame analysis pertain to the detection-based VSDM group. At the same time, only the work in Pouw et al. (2020) belongs to the trajectory-based VSDM category, which is based on quantifying the spatiotemporal trajectories distances to address the SD problem in a dynamic manner. Put simply, detection-based VSDM techniques aim to detect and calibrate individuals’ positions and then analyze the frames one by one to measure the distances between detected individuals in the BEV. By contrast, the trajectory-based VSDM approaches track people and calibrate the trajectories. Thereafter, the corresponding calibrated trajectories in the 3D spatiotemporal coordinates (in addition to the time axis) are used to determine distances between detected pedestrians. For better monitoring of the SD, continuous measurement and analysis over time is more appropriate rather than a specific moment. Therefore, more research focus should be put on investigating the VSDM problem based on analyzing spatiotemporal trajectories over time.

Besides, most of existing VSDM have reached excellent performance in terms of the accuracy of detecting pedestrians and measuring the distance between them in a low and medium density scene. This is due to the capability of cameras in easily tracking moving objects, and measuring physical distances in such environments. However, performing these task in crowded and dense places rests challenging. Indeed, some pedestrians can be hidden together, and hence they become invisible in crowded gatherings, ever for human observers. Consequently, it is quite difficult to put BBs to all pedestrians in dense scenes (Sahraoui et al., 2020).

6.3. Open challenges

6.3.1. Pedestrian overlapping and sensor noise

Pedestrian overlapping and occlusion is a serious problem that can considerably bias the results of VSDM systems. Also, the distance calculation can be inaccurate in some indoor applications due to the limited height and space. Typically, the need for video data from multiple cameras is significant. While this option can be achieved in both indoor and outdoor scenarios by installing numerous cameras and collecting different views, adopting drone-based surveillance that has the flexibility to move and monitor pedestrians can be another option for outdoor application scenarios.

Moreover, as most reviewed VSDM frameworks have relied on using ML and DL tools, the probability that a detected violation is a false alarm (due to the sensor noise or other reasons) has been assessed using different ML metrics, e.g., confusion matrix, false alarm rate (FAR), etc. For instance, in Rahim et al. (2021), a YOLOv4-based VSDM solution under different low light conditions is proposed, demonstrating good reliability to light changes (which can be considered as sensor noise). Typically, no single false-positive (FP) has been detected.

6.3.2. Computational complexity

Based on the literature review, some studies have successfully achieved real-time SD monitoring (including pedestrian detection, calculation of interpersonal distances between pedestrians, violation detection and generation of alerts) in moderately dense crowds, such as Chandel et al. (2020), Nakano and Nishimura (2021), Pouw et al. (2020), Sahraoui et al. (2020), Saponara et al. (2021), Saponara, Elhanashi, and Zheng (2022), Shao et al. (2021). Besides, other commercial solutions have also been developed for real-time monitoring of SD, e.g., dRISK (drisk: Real-time monitoring of social distancing, 2022), based on predicting distances between individuals using a single monocular CCTV camera. Similarly, the live SD monitoring (LSDM) solution (LSDM: live social-distancing monitoring solution, 2022), developed by Intel, achieves real-time tracking and monitoring of pedestrians using distributed computing, AI models, and radar sensors. Moreover, it enables the representation of detected pedestrians as live, contextual insights and reporting on web-based dashboards.

However, the complexity of VSDM systems (e.g., detecting all mutual distances) increases with the increased density of monitored crowds (i.e., the rise in the number of observed people). For instance, the computational complexity of the VSDM system introduced by Al-Sa’d et al. (2022) has been measured by its frame rate (the number of processed video frames per second) and processing rate (i.e., the amount of processing time per frame). This VSDM system includes (i) person detection and localization, (ii) top-view transformation, (iii) smoothing/tracking (smooth noisy top-view positions and compensate for missing data due to occlusion with tracking), (iv) distance measurement and (v) violation detection. This has been done for two cases, i.e., without and with the smoothing/tracking stage on a desktop equipped with 2 Intel Xeon E5-2697V2 x64-based processors and has 192 GB of memory. Fig. 13 portrays computational complexity analysis results obtained for both scenarios. The capability of the system to run in real-time has been demonstrated by the average results, although the smoothing/tracking stage can add more computational complexity. In this regard, the VSDM system can be run at 106.5 fps (9.9 ms/frame) without the smoothing/tracking stage, while that could be decreased to run at 33.6 fps (44.5 ms/frame) when accommodating the smoothing/tracking algorithm. On the other hand, it is also confirmed from the reported results that increasing the number of tracked people can significantly augment the computational cost. This leads to lower frame and processing rates in both scenarios (i.e., without and with the smoothing/tracking stage).

| Inter-pedestrian distance | Truth value (m) | Measured value (m) | Absolute error (m) | SD accuracy |
|---------------------------|----------------|--------------------|--------------------|-------------|
| 3~4                       | 1.980          | 1.905              | 0.075              | 0.962       |
| 4~5                       | 1.980          | 1.855              | 0.125              | 0.936       |
| 3~6                       | 1.980          | 1.877              | 0.103              | 0.947       |
| 5~6                       | 1.980          | 1.860              | 0.120              | 0.939       |
| 2~7                       | 1.980          | 1.897              | 0.083              | 0.958       |
| 4~7                       | 1.980          | 1.859              | 0.121              | 0.938       |
| 2~8                       | 1.980          | 1.855              | 0.125              | 0.936       |
| 4~8                       | 1.980          | 1.853              | 0.127              | 0.935       |
| Average                   | 1.980          | 1.870              | 0.109              | 0.945       |
Sustainable Cities and Society 85 (2022) 104064

Y. Himeur et al.

6.3.3. Camera calibration

Some VSDM frameworks have different limitations among them camera calibration which is performed manually (Nakano & Nishimura, 2021). Even worst, for some datasets, the floor plan or the transformation matrix are missing. Thus, the authors need to estimate the size of a reference object in video frames by comparing it with the width of detected pedestrians and then utilize the key points of the reference object to measure the perspective transformation. In this respect, a transformation can be produced and used for camera calibration. To overcome the problem of camera’s calibration, Nakano and Nishimura (2021) introduce a two-stage automatic VSDM, which is based on (i) camera auto-calibration (offline) using human joints to determine the 3D position and rotation of the camera, and (ii) pedestrian detection using pedestrian pose detection; (iii) pedestrian’s 3D location detection using estimated calibrated data, distance measurement in the BEV.

Most of drone-based and CCTV cameras are collecting tilted images, which makes challenging their transformation to real-world coordinates. However, in some research studies such as Dubrofsky (2009), a homography exists between video frames recorded with the same camera for the same area at different positions or angles. Typically, a homography exists between two planes of the same area that correspond to tilted and vertical images, respectively. A homography transformation can then be used to transform the tilted images into real-world coordinates. In this respect, solving this problem requires transforming the tilted images to the vertical images using a homography matrix, and transforming the vertical images to the real-world coordinates based on the concept of transforming vertical images to real-world coordinates. Fig. 14 portrays an example of calibrating tilted images, where $H$ is a homography matrix as defined in Dubrofsky (2009) and $\lambda$ refers to the ratio of pixel to meter.

6.3.4. Lack of annotated datasets

Because of the privacy concerns and lockdown measures set in various countries, producing large-scale datasets for validating VSDM monitoring solutions was challenging. To close this gap, virtual reality (VR) is used in Mukhopadhyay, Reddy, Ghosh, LRD, and Biswas (2021), Mukhopadhyay, Reddy, Saluja, et al. (2021), to generate customized datasets and validate DL-based VSDM algorithms. Typically, VR can provide capabilities of interaction between individuals in a shared 3D environment. This opens the doors for various shared activities and experiences that could not be possible with other remote communication modalities.

In this regard, VR has been adopted in Mukhopadhyay, Reddy, Ghosh, et al. (2021) to model a digital twin of an office space and utilized it to produce a comprehensive dataset of users in various locations, dresses and postures. Besides, in Mukhopadhyay, Reddy, Saluja, et al. (2021), a CNN-based VSDM system is implemented to detect individuals in a limited-sized dataset of real humans, which has been augmented with a simulated dataset of humanoid figures. Typically, the VR environment has been improved using an interactive dashboard, which shows information gathered from physical sensors and the latest statistics on COVID-19. On the other hand, YOLOv3 has been utilized to detect people in VR environments. Moving on, in Priyan, Johar, Alkawaz, and Helmi (2021), VR, smartphones, and IoT devices are used to monitor the compliance of pedestrians with SD norms. Specifically, this has been possible by visually enabling people to control their distances in the real-world based on their mobile cameras by using an augmented reality app.

6.3.5. Security and privacy concerns

VSDM techniques are based on mass surveillance of crowds and individuals in public areas, thus, it has been imperative to catch some potential impacts on the surrounding environments. Typically, an entire adherence to safety guidelines is not ensured as the VSDM technology is susceptible to human error and corrupt with different privacy breaches. Specifically, exchanging images/videos including information about detected individuals with data centers and responsive authorities to penalize SD violators can represent a serious privacy issue (Sugianto, Tjondronegoro, Stockdale, & Yuwono, 2021). Additionally, numerous complaints have been raised about increased panic and anxiety among the individuals receiving repetitive alerts.
To that end, developing systems that automate the SD monitoring procedure with high security and privacy preservation levels is becoming an urgent need. More, recently, few studies have been proposed to overcome these issues. For instance, in Al-Sa’d et al. (2022) a privacy-preserving VSDM method for CCTV cameras is proposed. Typically, a person localization method is developed based on pose estimation. Next, a privacy-preserving adaptive smoothing and tracking approach is built for (i) mitigating noisy/missing measurements and occlusions, (ii) computing distances between pedestrians (in the real-world coordinates), detecting SD violations, and identifying overcrowded areas in scenes. Moving on, CNN models and the blockchain technology have been leveraged in Tanwar et al. (2021) to monitor SD. If SD violations are detected the surveillance center is alerted via blockchain and necessary actions are then taken. Another solution to alleviate the privacy issues relies on adopting BEV cameras. In this regard, because of the privacy concerns raised when deploying street-level cameras to record videos, Ghasemi, Yang, et al. (2021) develop a BEV-based SD analyzer (B-SDA), which helps preserve pedestrians’ privacy by using BEV cameras.

6.3.6. Detection of family-groups and safe social groups (SSGs)

In most existing VSDM frameworks, SD violations are defined as instances where mutual distances between pairs of individuals become lower than a predefined threshold. Typically, these are considered as violations without any exceptions. However, this must not be the case for family-groups, as they are allowed to stay closer and no alerts should be triggered. To that end, it is important to discriminate between “safe social groups (SSGs)” and random pedestrians in close proximity to each other. A SSG can be defined as ensemble of individuals supposed to reside together, e.g., a family (Yang, Sun, et al., 2021).

Although the importance of this point, few studies have been investigated to detect family-groups or safe social groups while monitoring the distance between pedestrians. For instance in Pouw et al. (2020), the authors focus on real-time trajectory detection and individual group analysis by imposing thresholds on the distance-time contact patterns. Typically, mutual distances and contact times have been considered along with statistical observables as the radial distribution functions (RDFs). They have conveniently been utilized for quantifying average exposure times. Therefore, the automation of definitions of family-groups and characterization of statistical distributions of violations have been enabled. In this respect, family members have been identified as the persons that persistently remain closer than a specific threshold distance for adequately long time. On the other hand, this helps define SD violations as those related to distance violations of individuals that only inconsistently (i.e., occasionally) yield COVID-19 events infringing the minimal distance rules.

On the other hand, if children and parents walk side-by-side, the rule on SD must be ignored even if the physical distance between them is less than the SD norm. Aiming at accounting for this particular scenario, an exclusion approach for child/parent pedestrians, defined based on pedestrians’ height in China, is proposed in Niu et al. (2021). The authors approve the standard of free admission for children in most public places (e.g., malls, amusement parks, cinemas, tourist attractions, etc.) and select a reference height of 1.2 m for children. Meanwhile, based on the pertinent statistical data in Vischer (2008), the adult height reference has been selected as the average height of 1.715 m. In this context, pedestrians walking side-by-side with a height difference of more than 51.5 cm could be regarded as family members, and hence, the SD violations would be bypassed.

7. Future directions

We highlight in this section the future research perspectives although it was proved in Section YOLO-based methods have reached excellent performance, especially in simple scenes, in terms of the accuracy and reliability. Typically, there are still some performance issues with complex scenes in addition to other problems which are mainly related to privacy preservation, lack of annotated datasets, camera calibration, etc. We present in what follows the future directions that can overcome these issues:

7.1. VSDM on the edge

While CNN-based VSDM systems provide excellent accuracy to monitor SD, they can be deployed in different kinds of public spaces, such as shopping areas, airports, parks, industrial areas, etc., to slow the spread of the virus. However, their application scenarios present serious challenges to the underlying computing platforms. Specifically, small, low-cost, and energy efficient computing boards must be used to promote their implementation and enable the use of mobile surveillance while maintaining sufficient computing power and memory to run robust CNN algorithms at a lower latency. Moreover, preserving privacy of pedestrians detected in VSDM systems requires to process data edge devices, without transmitting it to cloud data centers (Fasfous et al., 2021). In this regard, exploring light-weight CNN algorithms and deploying them on edge and mobile devices can be the great option. This helps avoid the privacy concerns as person-specific data is processed on the edge/mobile device closer to the monitoring entity. This also aids in implementing VSDM system on drones to benefit from their flexibility. Although light-weight CNN models have a fast inference, their main challenge is their low pedestrian detection accuracy, especially in dense crowds (Quinonez & Torres, 2022; Restas, 2022).

As presented in Section 5.2.3, various lightweight CNN-based models have been introduced that are appropriate for mobile platforms, such as ShuffleNet, SqueezeNet, MobileNet (Kong et al., 2021). However, these models depend considerably on deep separable convolution and lack effective implementations in some DL frameworks. To that end, Shao et al. (2021) use PeleeNet, a light-weight CNN model perform real-time VSDM on images recorded from drones. The implementation of PeleeNet is completed using conventional convolution and features are extracted with fewer parameters. Similarly, a VSDM solution is developed by elnfochips (AI Vision Based Social Distancing Detection, 2022), which is powered by NVIDIA Jetson AGX Xavier (Jetson agx xavier developer kit, 2022). Typically, after decoding and preprocessing video recorded with, pre-trained DL algorithms detect pedestrians. Moving forward, insights about individual density/SD are extracted in real-time. Extracted information is locally saved on edge devices then moved to a cloud platform, which is only accessible to security managers or concerned authorities for (i) reviewing SD, and (ii) taking appropriate actions in case of any violations. Besides, in Ramadass et al. (2020), a OLOv3-based VSDM is embedded in a drone’s camera, which runs the yolov3 algorithm and detects if the SD is respected or not and if people are wearing masks.

7.2. Federated learning (FL)

VSDM, as a computer vision-based DL technology, requires saving video data on cloud platforms for centralized training (especially for pedestrian detection and tracking). However, this cannot be the best methodology because of the high cost of transmitting video data and privacy concerns. Accordingly, as presented in Table 1, many frameworks included in this review have failed to address the privacy concerns. FL has recently been introduced to separate the requirement of powerful DL from the need to store large-scale datasets in the clouds. Specifically, FL is a distributed ML technique that relies on the storage and computing capacity of the devices themselves (e.g., cameras) to co-build DL models without transferring data to the cloud, hence, without adversely affecting the privacy of individuals (Zhu, Yin, Xiong, Tang, & Yin, 2021).
Sustainable Cities and Society 85 (2022) 104064

Y. Himeur et al.

Fig. 15. mAP vs. gpu computation cost for (a) CNN-based object detectors, and (b) CNN-based feature extractors. Every object detector (or feature extractor) can correspond to different on the graph because of the changing input strides, sizes, etc. (Huang et al., 2017).

7.3. Deep transfer learning (DTL) for better generalization

Developing DTL- and deep domain adaptation (DDA)-based VSDM schemes will help increase the generalization of these algorithms on datasets with entirely different characteristics. Typically, YOLOv3, Yolov4, and Faster-RCNN have successfully been applied to different simple datasets; however, using them on other complex datasets is still challenging. Also, processing different datasets with distinct image resolutions is still challenging as some DL models process image inputs with a fixed size of images. Thus, resizing these images is required, although this generally results in information deficiency and object distortion, which can be a possible restriction. Accordingly, applying DDA or DTL for processing numerous image resolutions considers a promising research direction to automate the VSDM task.

Up to now, all existing methods have used DTL for the pedestrian detection task. However, for better SD monitoring, it is worth applying DTL and DA to predict the physical distances among pedestrians. This is doable by first developing automatic and labeled SD monitoring datasets based on a rendering engine simulation (Di Benedetto et al., 2022).

7.4. Real-time VSDM

Implementing a real-time VSDM system requires optimizing two primary parameters, i.e., the accuracy of pedestrian detection and computation cost. Typically, the first parameter is usually represented by the mAP while the second one refers to computation time or the number of processed frames per second (fps). In this respect, the performance of the VSDM system will be increased if the computation cost is low and the mAP is high. Fig. 15 portrays a scatter plot of the mAP vs. the GPU computation times for different CNN-based object detectors (meta-architectures) and CNN-based feature extractors (Huang et al., 2017).

To enable the real-time operation of VSDM systems, all the stages involved in the SD monitoring should be run in real-time, including pedestrian detection and tracking, interpersonal distance estimation, detection of violations, and alert generation. For pedestrian detection and tracking, there has been a significant amount of studies targeting this real-time implementation since this topic has attracted substantial complexity issue that occurred due to (i) the increasing crowd density and (ii) using complex pedestrian detectors with large batch sizes.

8. Conclusion

This paper presented, to the best of the authors’ knowledge, the first comprehensive review of recent advances in the field of VSDM. In doing so, we first introduced the background of the VSDM problem after describing the survey methodology and explaining the article selection approach. Thereafter, evaluation metrics used in the overviewed articles are briefly presented. Next, the surveillance methodology employed to perform the VSDM, including fixed and drone-based, is explained.

Moving forward, existing VSDM contributions were discussed after categorizing them into two groups: techniques based on hand-crafted features and CNN-based methods. CNN-based methods have been classified into two categories with reference to the number of processing stages: single-stage and two-stage schemes. Also, these approaches have been classified into two main categories corresponding to the complexity of the CNN models used in each framework: complex and lightweight models. Additionally, the results of representative techniques are summarized according to the original literatures, and their pros and cons were identified. Overall, YOLOv3-based methods were the mainstream and promising techniques as up to 99.8% accuracy has been reached. However, the performance of existing methods is relative since they have been validated on different datasets. Thus, it is still challenging to conduct a fair comparison. Moreover, most existing VSDM frameworks have been tested in low- or medium-density scenes. By contrast, different areas in smart cities suffer from dense
and crowded gatherings, particularly at peak periods, which makes monitoring SD between pedestrians a challenge. While VSDM techniques can smoothly detect and track pedestrians and calculate physical distances in low or medium crowded scenes, they have some difficulties performing well in highly dense areas. Concretely, some pedestrians can be hidden together and become not visible in crowded gatherings, even to human observers.

All in all, although the intense attention paid by the research community to develop VSDM solutions in the hopes of combating the COVID-19 pandemic, the critical analysis enabled identified various open challenges, such as pedestrian overlapping, cameras’ calibration, lack of annotated datasets, and security and privacy concerns. Therefore, future directions that help overcome these issues and attract considerable research and development in the near future have been highlighted, including moving the VSDM algorithms to edge and mobile devices, using federated learning to promote privacy preservation, and adopting DTL for a better generalization of existing algorithms.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This research work was made possible by research grant support (QUEX-CENG-SCDL-19/20-1) from Supreme Committee for Delivery and Legacy (SC) in Qatar. The statements made herein are solely the responsibility of the authors. Open Access funding provided by the Qatar National Library.

References

Agarwal, N., Meena, C. S., Raj, B. P., Saini, L., Kumar, A., Gopalakrishnan, N., et al. (2021). Indoor air quality improvement in COVID-19 pandemic. Sustainable Cities and Society, 70, Article 102942.

agendran, A., Harper, D., & Shah, M. (2021). Ucf-arg dataset. University of Central Florida, URL http://crcv.ucf.edu/data/UCF-ARG.php.

Ahmed, I., Ahmad, M., & Jeon, G. (2021). Social distance monitoring framework using deep learning architecture to control infection transmission of COVID-19 pandemic. Sustainable Cities and Society, 70, Article 102777.

Ahmed, I., Ahmad, M., Rodrigues, J. J., Jeon, G., & Din, S. (2021). A deep learning-based social distance monitoring framework for COVID-19. Sustainable Cities and Society, 65, Article 102571.

Ahmed, I., Jeon, G., Chethri, A., & Hassan, M. M. (2021). Adapting Gaussian YOLOv3 with transfer learning for overhead view human detection in smart cities and societies. Sustainable Cities and Society, 70, Article 102908.

AI vision based social distancing detection. (2022). https://www.einfochips.com/resources/success-stories/ai-vision-based-social-distancing-detection/. (Accessed 13 January 2022).

Al-Sa’id, M., Kiranayaz, S., Ahmad, I., Sundell, C., Vakkuri, M., & Gabbouj, M. (2022). A social distance estimation and crowd monitoring system for surveillance cameras. Sensors, 22(2), 418.

Amato, G., Ciampi, L., Falchi, F., Gennaro, C., & Messina, N. (2019). Learning pedestrian detection from virtual worlds. In International conference on image analysis and processing (pp. 302–312). Springer.

Anitha Kumari, K., Purusothaman, F., Dharni, D., & Padmashani, R. (2021). COVID-19: An enabled social distancing detector using CNN. In Computational intelligence techniques for combating COVID-19 (pp. 95–115). Springer.

Amari, M., Singh, D. K., et al. (2021). Monitoring social distancing through human detection for preventing/reducing COVID spread. International Journal of Information Technology, 13(3), 1255–1264.

Ar, M. L. A., Nugraha, Y., Ernesto, A., Kanggrawan, J. I., Suherman, A. L., et al. (2020). A computer vision-based object detection and counting for COVID-19 protocol compliance: A case study of Jakarta. In 2020 International conference on ICT for smart society (pp. 1–5). IEEE.

Benfolds, B., & Reid, I. (2011). Stable multi-target tracking in real-time surveillance video. In CVPR 2011 (pp. 3457–3464). IEEE.

Bierke, A., Bakker, M., Vepakomma, P., Raskar, R., Larson, K., & Pentland, A. (2020). Assessing disease exposure risk with location histories and protecting privacy: A cryptographic approach in response to a global pandemic. arXiv preprint arXiv:2010.14412.

Bertoni, L., Kreiss, S., & Alahi, A. (2019). Monoloco: Monocular 3d pedestrian localization and uncertainty estimation. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 6861–6871).

Bertoni, L., Kreiss, S., & Alahi, A. (2021). Perceiving humans: From monocular 3D localization to social distancing. IEEE Transactions on Intelligent Transportation Systems.

Bewley, A., Ge, Z., Ott, L., Ramos, F., & Upcroft, B. (2016). Simple online and realtime tracking. In 2016 IEEE international conference on image processing (pp. 3464–3468). IEEE.

Bian, S., Zhou, B., Bello, H., & Lukowicz, P. (2020). A wearable magnetic field based proximity sensing system for monitoring COVID-19 social distancing. In Proceedings of the 2020 international symposium on wearable computers (pp. 22–26).

Bianco, G. M., Panunzio, N., & Marrocchi, G. (2021). RFID research against COVID-19—sensorized face masks. In 2021 IEEE international conference on RFID technology and applications (pp. 241–243). IEEE.

Borra, S. (2020). COVID-19 apps: Privacy and security concerns. In Intelligent systems and methods to combat Covid-19 (pp. 11–17). Springer.

Boubel, F., Milki, H., & Hammami, M. (2021). Crowd behavior analysis based on convolutional neural network: Social distancing control COVID-19. In VISIGRAPP (pp. 273–280).

Braithwaite, I., Gallender, T., Bullock, M., & Aldridge, R. W. (2020). Automated and partly automated contact tracing: A systematic review to inform the control of COVID-19. The Lancet Digital Health.

Chandel, V., Banerjee, S., & Ghose, A. (2020). Proxirank: A robust solution to enforce real-time social distancing & contact tracing in enterprise scenario. In Adjunct proceedings of the 2020 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2020 ACM international conference on wearable computers (pp. 503–511).

Chaudhary, K. (2020). Maintaining social distancing using artificial intelligence. Available at SSRN 3668540.

Chavarrova, T., Baqiet, P., Bouquet, S., Makasi, A., Jons, C., Bagautdinov, T., et al. (2018). Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5030–5039).

Chen, K., Loy, C. C., Gong, S., & Xiang, T. (2012). Feature mining for localised crowd counting. In BMVC, vol. 1, no. 2 (p. 3).

Cho, H., Ippoliti, D., & Yu, W. Y. (2020). Contact tracing mobile apps for COVID-19: Privacy considerations and related trade-offs. arXiv preprint arXiv:2003.11511.

Conte, C., de Alteris, G., De Pandi, F., Caputo, E., Moriello, R. S. L., Rufino, G., et al. (2021). Performance analysis for human crowd monitoring to control COVID-19 disease by drone surveillance. In 2021 IEEE 8th international workshop on metrology for aerospace (pp. 31–36). IEEE.

Cristiani, M., Del Bue, A., Murino, V., Setti, F., & Vinciarelli, A. (2020). The visual social distancing problem. IEEE Access, 8, 126876–126886.

Dai, Z., Jiang, Y., Li, Y., Liu, B., Chan, A. B., & Vasconcelos, N. (2021). BEV-net: Assessing social distancing compliance by joint people localization and geometric reasoning. In Proceedings of the IEEE/CVF international conference on computer Vision (pp. 5401–5411).

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., et al. (2017). Deformable convolutional networks. In Proceedings of the IEEE international conference on computer vision (pp. 764–773).

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In 2005 IEEE computer society conference on computer vision and pattern recognition, vol. 1 (pp. 866–873). Ieee.

Datakalab [analyse de l’image par ordinateur. (2022). https://www.datakalab.com/?lang=en. (Accessed 31 January 2022).

Davis, J. W., & Sharma, V. (2007). Background-subtraction using contour-based fusion with learning. In Proceedings of the 2005 IEEE computer society conference on computer vision and pattern recognition, vol. 2 (pp. 886–893). Ieee.

Dey, S. (2019). A computer vision-based object detection and counting for COVID-19 protocol. In Proceedings of the 2019 international symposium on metrology and methods to combat Covid-19 (pp. 11–17). Springer.

Degadwala, S., Vyas, D., Dave, H., & Mahajan, A. (2020). Visual social distance alert system using computer vision & deep learning. In 2020 4th international conference on electronics, communication and aerospace technology (pp. 1512–1516). IEEE.

Del Benedito, M., Carrara, F., Ciampi, L., Falchi, F., Gennaro, C., & Amato, G. (2022). An embedded toolkit for human activity monitoring in critical environments. Expert Systems with Applications, Article 117125.

DRISK: Real-time monitoring of social distancing. (2022). https://drisk.ai/real-time-monitoring-of-social-distancing. (Accessed 21 June 2022).

Du, X., Song, Y., Yang, B., & Zhou, Y. (2022). Strongsort: Make deepsort great again. arXiv preprint arXiv:2202.13514.
Madane, S., & Chitre, D. (2021). Social distancing detection and analysis through computer vision. In 2021 6th International Conference for Convergence in Technology (pp. 1–10). IEEE.

Magon, R., Singh, H., Jindal, N., Hooda, N., & Rana, P. S. (2021). Deep learning-based bird eye view social distancing monitoring using surveillance video for curbing the COVID-19 spread. Neural Computing and Applications, 33(22), 15867–15814.

Manzira, C. K., Charly, A., & Caulfield, B. (2022). Assessing the impact of mobility on the incidence of COVID-19 in Dublin city. Sustainable Cities and Society, 80, Article 102770.

Mbugue, E. (2020). Integrating emerging technologies into COVID-19 contact tracing: Opportunities, challenges and pitfalls. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 14(6), 1631–1636.

Meivel, S., Siddhwan, N., Anand, R., Pandey, D., Alnuaim, A. A., Atheneeyan, A. S., et al. (2022). Mask detection and social distance identification using internet of things and faster R-CNN algorithm. Computational Intelligence and Neuroscience, 2022.

Mercaldo, F., Martellini, F., & Santone, A. (2021). A proposal to ensure social distancing with deep learning-based object detection. In 2021 International Joint conference on neural networks (pp. 1–5). IEEE.

Meynberg, O., & Kuschki, G. (2013). Airborne crowd density estimation. ESFR Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2, 49–54.

Milan, A., Leaf-Taixé, L., Reid, I., Roth, S., & Schindler, K. (2016). MOT16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.08831.

Mohamed, S. K., & Abdel Samee, B. E. (2022). Social distancing model utilizing machine learning techniques. In Advances in data science and intelligent data communications technologies for COVID-19 (pp. 41–53). Springer.

Mukhopadhyay, A., Reddy, G. R., Ghosh, S. R., B. M., & Biswas, P. (2021). Validating social distancing through deep learning and VR-based digital twins. In Proceedings of the 27th ACM symposium on virtual reality software and technology (pp. 1–2).

Mukhopadhyay, A., Reddy, G. R., Ghosh, S. R., Petro-Biao, A. G., & et al. (2021). Virtual-reality-based digital twin of office spaces with social distance measurement feature. Virtual Reality & Intelligent Hardware, XXXX, XX (XX), 1–21.

Nagrath, P., Jain, R., Madan, A., Arora, R., Kataria, P., & Hemanth, J. (2021). SMDNMV2: A real-time DNN-based face mask detection system using single shot multibox detector and MobileNetV2. Sustainable Cities and Society, 66, Article 102692.

Nakano, G., & Nishimura, S. (2021). Real-time social distancing detection system with auto-calibration using pose information. In The first International conference on AI-MLS (pp. 1–3).

Nguyen, N. T., Tap, V. M., Van Huyhn, N., Nguyen, N. T., Khoa, T. V., Tuan, B. M., et al. (2020). A comprehensive survey of enabling and emerging technologies for social distancing—Part I: Fundamentals and enabling technologies. IEEE Access, 8, 153479–153507.

Niu, Y., Xu, Z., Xu, E., Li, G., Huo, Y., & Sun, W. (2021). Monocular 3D localization for social distance monitoring. Sensors, 21(17), 5908.

Oh, S., Hsiao, A., Pereira, A., Gunturk, N., Chen, C.-C., Lee, J. T., et al. (2011). A large-scale benchmark dataset for event recognition in surveillance video. In CVPR 2011 (pp. 3153–3160). IEEE.

Oranrirkul, T., & Takada, H. (2020). Social distancing warning system at public transportation using Wi-Fi signal from mobile devices. In Advances of the 2020 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2020 ACM international symposium on wearable computers (pp. 267–271).

Özbek, M. M., Syed, M., & Öksüz, I. (2021). Subjective analysis of social distance violation detection. IEEE Access, 9, 159915–159930.

Poonam, N. S., Sonbhadra, S. K., Agarwal, S., & Rai, G. (2020). Monitoring COVID-19 spread. Journal of Physics: Conference Series, 153479–153507.

Qin, J., & Xu, X. (2021). Research and implementation of social distancing monitoring technology based on SSD. Procedia Computer Science, 183, 768–775.

Quintonez, F., & Torres, R. (2022). Evaluation of ai0 performance in cloud and edge for deep convolutional models for mask detection. Image and Vision Computing, 86, 104064.

Ramadass, L., Arunachalam, S., & Sayagares, Z. (2020). Applying deep learning algorithm to maintain social distance in public place through drone technology. International Journal of Pervasive Computing and Communications.

Ramachandani, A., Fan, C., & Mostafavi, A. (2020). Deepvivid: An interpretable deep learning model for predictive surveillance of covid-19 using heterogeneous features and their interactions. IEEE Access, 8, 159915–159930.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779–788).

Redmon, J., & Farhadi, A. (2017). YOLO9000: Better, faster, stronger. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7263–7271).

Ren, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems, 28.

Resaie, M., & Arazmi, M. (2020). Deep social: Social distancing monitoring and infection risk assessment in covid-19 pandemic. Applied Sciences, 10(21), 7514.

Ramadass, L., Arunachalam, S., & Sagayasree, Z. (2020). Applying deep learning for social distancing monitoring of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 22–26.

Rahim, A., Masghoul, A., Mirza, A., Afzal, F., & Anghar, I. (2022). DeptSel: An improved deep-learning and time-of-flight-based real-time social distance monitoring approach under various low-light conditions. Electronics, 11(3), 458.

Rahim, A., Masghoul, A., & Rana, T. (2021). Monitoring social distancing under various low-light conditions with deep learning and a single monochrome time of flight camera. PLoS One, 16(2), Article e0247440.

Rahim, A., & Mostafavi, A. (2020). Monitoring social distancing constraints in crowded scenarios. arXiv preprint arXiv:2008.06585.

Shah, J., Chaudhary, M., & Kanan, P. (2021). Social distancing detection using computer vision. In 2021 5th International Conference on computing methodologies and applications (pp. 1359–1366). http://dx.doi.org/10.1109/ICCM51019.2021.9418312.

Shah, J., Chaudhary, M., & Kanan, P. (2021). Social distancing detection using computer vision. In 2021 5th International Conference on computing methodologies and applications (pp. 1359–1366). http://dx.doi.org/10.1109/ICCM51019.2021.9418312.

Shalini, G., Marget, M. K., Niraimathi, M. S., & Subbarah, S. (2021). Social distancing analyzer using computer vision and deep learning. Journal of Physics: Conference Series, 1916(1), Article 012039.

Shao, Z., Cheng, G., Ma, J., Wang, Z., Wang, J., & Li, D. (2021). Real-time and accurate UAV pedestrian detection for social distancing monitoring in COVID-19 pandemic. IEEE Transactions on Multimedia.

Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., et al. (2018). Crowdhuman: A large-scale benchmark for detecting human in a crowd. arXiv preprint arXiv:1805.00123.

Sharan, J., Chanu, N. I., Jena, A. K., Arunachalam, S., & Choudhary, P. K. (2020). Social distancing detection and analysis through computer vision. In 2020 12th Control and System Graduate Research Conference (CGRCON) (pp. 7263–7271).

Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., et al. (2018). Crowdhuman: A large-scale benchmark for detecting human in a crowd. arXiv preprint arXiv:1805.00123.
Shrestha, S., Lu, D., Tian, H., Cao, Q., Liu, J., Rizzo, J.-R., et al. (2020). Active crowd analysis for pandemic risk mitigation for blind or visually impaired persons. In European conference on computer vision (pp. 422–439). Springer.

Srivastava, S., Zhao, X., Manay, A., & Chen, Q. (2021). Effective ventilation and air disinfection system for reducing coronavirus disease 2019 (COVID-19) infection risk in office buildings. Sustainable Cities and Society, 75, Article 103408.

Su, J., He, X., Qinq, L., Niu, T., Cheng, Y., & Peng, Y. (2021). A novel social distancing analysis in urban public space: A new online spatio-temporal trajectory approach. Sustainable Cities and Society, 68, Article 102765.

Sugianto, N., Tjondronegoro, D., Stockdale, R., & Yuwono, E. I. (2021). Privacy-preserving AI-enabled video surveillance for social distancing: Responsible design and deployment for public spaces. Information Technology & People.

Tang, F., Feng, Y., Chibeh, H., & Fan, J. (2021). The interplay of demographic variables and social distancing scores in deep prediction of US COVID-19 cases. Journal of the American Statistical Association, 116(534), 492–506.

Tanwar, S., Gupta, R., Patel, M. M., Shukla, A., Sharma, G., & Davidson, I. E. (2021). Blockchain and AI-enabled social distancing scheme to combat COVID-19 situations. IEEE Access, 9, 129830–129840. http://dx.doi.org/10.1109/ACCESS.2021.3114098.

Teboulbi, S., Messaoud, S., Hajjaji, M. A., & Mitha, A. (2021). Real-time implementation of AI-based face mask detection and social distancing measuring system for COVID-19 prevention. Scientific Programming, 2021.

To, H.-T., Bui, K.-H. N., Le, V.-D., Bui, T.-C., Li, W.-S., & Cha, S. K. (2021). Real-time social distancing alert system using pose estimation on smart edge devices. In International conference on intelligent information and database systems (pp. 291–300). Springer.

Tomáš, J., Rego, A., Viccians-Tudela, S., & Lloret, J. (2021). Incorrect facemask-wearing detection using convolutional neural networks with transfer learning. In Healthcare, Vol. 9, no. 8 (p. 1050). Multidisciplinary Digital Publishing Institute.

Udugama, B., Kadhiresan, P., Kozlowski, H. N., Malekjahani, A., Osborne, M., Li, V. Y., et al. (2020). Diagnosing COVID-19: The disease and tools for detection. ACS Nano, 14(4), 3822–3835.

Using 3D cameras to monitor social distancing. (2022). https://www.stereolabs.com/blog/using-3d-cameras-to-monitor-social-distancing/. (Accessed 31 January 2022).

Usman, M., Lee, T.-C., Moghe, R., Zhang, X., Faloutsos, P., & Kapadia, M. (2020). A social distancing index: Evaluating navigational policies on human proximity using crowd simulations. In Motion, interaction and games (pp. 1–6).

Valencia, I. J. C., Dadios, E. P., Fillone, A. M., Puno, J. C. V., Baldovino, R. G., & Billones, R. K. C. (2021). Vision-based crowd counting and social distancing monitoring using tiny-YOLOv4 and DeepSORT. In 2021 IEEE International smart cities conference (pp. 1–7). IEEE.

Varghese, E. B., & Thampi, S. M. (2021). A multimodal deep fusion graph framework to detect social distancing violations and FCOs in pandemic surveillance. Engineering Applications of Artificial Intelligence, 103, Article 104305.

Visscher, P. M. (2008). Sizing up human height variation. Nature Genetics, 40(5), 489–490.

Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. (2021). Scaled-yolov4: Scaling cross stage partial network. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 13029–13038).

Wang, L., Shi, J., Song, G., Shen, I.-F., et al. (2007). Object detection combining recognition and segmentation. In Asian conference on computer vision (pp. 189–199). Springer.

Who coronavirus (COVID-19) dashboard. (2022). https://covid19.who.int/. (Accessed 13 January 2022).

Widiatmoko, F., Berchmans, H. J., & Setiawan, W. (2021). Computer vision and deep learning approach for social distancing detection during Covid-19 pandemic (Ph.D. thesis), Swiss German University.

Wojke, N., Bewley, A., & Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. In 2017 IEEE international conference on image processing (pp. 3645–3649). IEEE.

Yang, Z., Sun, M., Ye, H., Xiong, Z., Zusman, G., & Kostic, Z. (2021). Birds eye view social distancing analysis system. arXiv preprint arXiv:2112.07159.

Yang, D., Yurtsever, E., Renganathan, V., Redmill, K. A., & Özgüner, U. (2021). A vision-based social distancing and critical density detection system for COVID-19. Sensors, 21(13), 4608.

Zhang, N., Su, B., Chan, P.-T., Miao, T., Wang, P., & Li, Y. (2020). Infection spread and high-resolution detection of close contact behaviors. International Journal of Environmental Research and Public Health, 17(4), 1445.

Zhang, Y., Wang, C., Wang, X., Zeng, W., & Liu, W. (2021). Fairmot: On the fairness of detection and re-identification in multiple object tracking. International Journal of Computer Vision, 129(11), 3069–3087.

Zhang, S., Wen, L., Bian, X., Lei, Z., & Li, S. Z. (2018). Single-shot refinement neural network for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4203–4212).

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., & Tian, Q. (2015). Scalable person re-identification: A benchmark. In Proceedings of the IEEE international conference on computer vision (pp. 1116–1124).

Zhou, B., Wang, X., & Tang, X. (2012). Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents. In 2012 IEEE conference on computer vision and pattern recognition (pp. 2871–2878). IEEE.

Zhu, X., Hu, H., Lin, S., & Dai, J. (2019). Deformable convnets v2: More deformable, better results. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9308–9316).

Zhu, R., Yin, K., Xiong, H., Yang, H., & Yin, G. (2021). Masked face detection algorithm in the dense crowd based on federated learning. Wireless Communications and Mobile Computing, 2021.

Ziran, H., & Dahnoun, N. (2021). A contactless solution for monitoring social distancing: A stereo vision enabled real-time human distance measuring system. In 2021 10th Mediterranean conference on embedded computing (pp. 1–6). IEEE.

Zoo, F., Gao, J., Kurkou, A., Yang, H., Orbay, K., & Ma, Q. (2021). Reference-free video-to-real distance approximation-based urban social distancing analytics amid COVID-19 pandemic. Journal of Transport & Health, 21, Article 101032.