Khan, W, Hussain, A, Muhammad Khan, B and Crockett, K

Outdoor mobility aid for People with visual impairment: Obstacle detection and responsive framework for the scene perception during the outdoor mobility of people with visual impairment

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Outdoor mobility aid for people with visual impairment: Obstacle detection and responsive framework for the scene perception during the outdoor mobility of people with visual impairment

Wasiq Khan [a,*], Abir Hussain [b], Bilal Muhammad Khan [c], Keeley Crockett [d]

[a] School of Computer Science and Mathematics, Liverpool John Moores University, Liverpool, Byrom Street L3 3AF, UK
[b] School of Engineering, University of Sharjah, Sharjah, UAE
[c] Department of Computer Science & Engineering, California State University, San Bernardino, CA 92407, USA
[d] Department of Computing & Mathematics, Manchester Metropolitan University, Manchester, UK

ARTICLE INFO

Keywords:
Scene perception
Assistive technology
Visual impairment mobility aid
Object detection dataset
Scene semantic representation
Responsive IoT devices

ABSTRACT

Outdoor mobility of individuals with visual impairment is challenging particularly where collision with obstacles can have significant impact on both physical and mental health. A variety of technological mobility aids for visually impaired people (VIP) have been studied and proposed in the literature which mainly utilise machine intelligence and deep learning (DL) approaches for object detection. However, object detection via the existing approaches suffers from reliability challenge due to real-time dynamics or the lack of available domain knowledge for specific obstacles identified by the VIP as potential hazards. In the present study, an object detection model (ObDtM) based on deep transfer learning techniques was developed for a custom-built dataset comprising of specific obstacles identified by the VIP as potential hazards. A custom dataset was compiled and manually annotated from various publicly available sources to train the ObDtM. Experiments were conducted to evaluate the proposed ObDtM for unseen obstacles kept as the test set. Results showed that ObDtM outperformed the state-of-the-art with 97% mean Average Precision (mAP), indicating a robust and generalizable DL approach. The compiled dataset and the ObDtM is useful for several potential use cases, particularly highlighting the use of DL in IoT and smart city applications. Additionally, a smart synergetic outdoor mobility framework was proposed for VIP (SOMAVIP) allowing comprehensive and accurate semantic representation of the surroundings by utilising the proposed ObDtM, cloud services, internet of things (IoT), and digital environment in the context of emerging smart city infrastructure. The proposed SOMAVIP can be highly impactful for improving VIPs' quality of life mainly for safer, cost-effective, and reliable independent outdoor mobility enriched with real-time perception and interpretations of the surroundings.

1. Introduction

The World Health Organisation (WHO) reported over 2.2 billion individuals with a vision impairment of some description across the globe (WHO, 2021) where, approximately 285 million are living with complete blindness or moderate to severe distance vision impairment (Ackland, Resnikoff, & Bourne, 2017). Visual impairment is the decreased ability to see, which poses a significant impact on the quality of life in a variety of aspects. Outdoor mobility (ODM) is of a major concern for VIP with a multitude of impacts including physical, social, and mental health. Independent mobility around local neighbourhood can be useful for social interactions and performing other daily activities (Wilson, 2015). However, various challenges are associated, particularly the unavailability of reliable assistive technology and mobility aids with their excessive cost. For instance, a service guide dog has an associated annual cost of approximately $48000, which is unaffordable for most of the VIP across the globe (Khan, Hussain, Khan, Nawaz, & Baker, 2019). Likewise, highly dynamic material and environmental factors (e.g., moving objects, image size, orientations, and varying backgrounds) pose a significant challenge to the reliability of existing autonomous assistive technologies for outdoor mobility of VIP (ODOMOVIP) (Khan, Hussain, Khan, Nawaz, & Baker, 2019).

[*] Corresponding author.
E-mail addresses: w.khan@ljmu.ac.uk (W. Khan), Abir.hussain@sharjah.ac.ae (A. Hussain), bilal.khan@csusb.edu (B.M. Khan), K.Crockett@mmu.ac.uk (K. Crockett).

https://doi.org/10.1016/j.eswa.2023.120464
Received 19 August 2022; Received in revised form 9 May 2023; Accepted 9 May 2023
Available online 12 May 2023

In 2015, a survey was conducted by the Royal National Institute of Blind People (RNIB) (Wilson, 2015) including approximately 500 VIP for whom a collision with an obstacle over a three-month period was reported. Among the participants, over 90% reported collisions were with street obstacles while walking in local surroundings and 33% reported injuries were reported to be (directly or indirectly) leading to increased costs such as medications, injury claims, and treatments. Majority of the VIP reported collisions were with parked cars (70%), recycling bins (64%), fixed street furniture (e.g., benches (60%), and advertising boards (49%)). Road crossings in local areas was also identified 'unsafe' (67%) by the VIP during ODM. Compared with fixed obstacles (e.g., street furniture), dynamic and temporary objects pose further challenges for VIPs' mobility in local streets. Around 50% of VIPs identified local roadworks to be seriously problematic for their outdoor mobility. Furthermore, cyclists using footpaths and temporary objects such as advertisement boards were reported as a "nightmare" for the ODOMOVIP and therefore, significantly affecting the VIPs quality of life in many aspects.

Recent digital transformation has led to an increased utilisation of smart data driven technologies in various aspects of human lives, including healthcare systems and assistive technologies. In addition to the traditional methods of VIPs' mobility support (e.g., guide dogs, human assistance), research and development have proposed various tools and techniques for both indoor and outdoor mobility of VIP within unfamiliar environments. For instance, various recent examples of related tools have been introduced that include the use of an infrared cane (Al-Fahoum, Al-Hmoud, & Al-Fraihat, 2013), voice and audio navigation (Nada, Fakhr, & Seddik, July 2015) (Simoes. & Lucena, 2016) (Kunta, C. Tuniki, & Sairam, 2020), laser (Wachaja & Agarwal, 2014, pp. 13–14) and ultrasonic sensors (Froneman & Heever, 2017), Microsoft Kinect and optical marker tracking (Zöllner, Huber, Jetter, & Reiterer, 2011), robot assistance (Capi & Toda, 2012), GPS navigation system with inertial measurement unit (IMU) (Zegarra & Farcy, 2012), RFID guiding cane (Liao, et al., 2013), LED lights and geomagnetic correction method (Nakajima & S. h, 2013), wearable RGBD camera (Lee. & Medioni, 2015), tactile perception with ultrasonic sensors (Ni, D, Song, A, Tian, L, Xu, X, & Chen, D, 2015), sonification of U-depth map of the surroundings (Skulimowski., P., Owczarek., M., Radecki., A., Bujacz., M., Rzeszotarski., D., & Strumillo, P., 2019), and Kinect depth camera (Ali, 2017). Furthermore, several methods have also been proposed based on data processing for the ODOMOVIP such as (Hoang, Nguyen, & Le, 2017) (Duarte, 2014) (Kumar., Y., & al., e., 2010) with limited application for indoor or simulated environments.

Despite the advancement for the available tools for the ODOMOVIP, generalisation, reliability, and intelligibility require are still in need of noticeable improvement. Additionally, increased deployments of the Internet Of Things (IoT) and responsive devices within smart applications and their integration with smart city concepts, are considered to be of significant influence for the reliability and performance of existing assistive technologies. For instance, e-Scooters can be considered a raising issue for VIP mobility independence mainly due to their noiseless features. Several news articles (Aspirot, 2021) including RNIB (Rnib, 2020) have highlighted e-Scooters as a 'nightmare' for VIP mobility independence. The RNIB recently reported (Rnib, 2020) that 81% of the VIP respondents preferred independent street walking, with e-Scooters viewed to be a major challenge to their mobility due to noiseless features. Even the latest assistive technologies are impractical for this kind of smart technological developments (Rnib, 2020) which indicates a dire need for a smart data driven technological solution that would fill the gap of existing tools in relation to the utilisation of IoT within the evolving digital environments and smart city applications.

Considering the limitations of existing mobility aids and assistive tools for ODOMOVIP, a smart data driven mobility aid was proposed for VIP (SOMAVIP), enabling real-time scene perception. The ODOMOVIP comprises of responsive devices and IoT embedded smart city environments, deep machine learning, computer vision, and data processing methods. Additionally, the present study proposed a real-time ObDtM for specific obstacles identified by the VIP within the RNIB survey (Wilson, 2015). The contributions of this work are as follows:

a) Annotated dataset (~60,000 images) for 8 objects of interest (with dynamic backgrounds) compiled from various publicly available sources. The dataset comprises diverse properties such as varying backgrounds, image size, resolution, orientation, and number of objects, which can be useful for generalising the proposed ObDtM as well as validating the object detection models used in this study in similar domains.
b) Multiple custom-trained ObDtM models were built for the ODOMOVIP utilising state-of-the art object detection techniques based on deep transfer learning. Custom trained ObDtM were then validated for diverse datasets which have also been made available to use for transfer learning in similar application domains.
c) A novel algorithm was developed for the proposed SOMAVIP framework for real-time scene interpretation for improving the perception of the surroundings based on the locality of VIP and detected objects. The proposed framework is aimed at supporting independent outdoor mobility for VIP by providing an enriched interaction with responsive IoT devices specifically, within the smart city environments.

The remainder of this manuscript is organised as follows. Section 2 addresses related works for the ODOMOVIP. Section 3 presents the proposed methodology for object detection using data annotations and deep transfer learning algorithms. Detailed statistical results from multiple models and discussion on the outcomes are provided in Section 4. Proposed SOMAVIP is presented in Section 5. Finally, the conclusion drawn from the proposed study are summarised in Section 6.

2. Related works

Assistive technologies for a VIPs' mobility can typically be categorised into outdoor (Jaime Sanchez, 2011) (Sanchez, 2008), indoor (Hub, Hartter, & Ertl, 2006) (Pinedo & Villanueva., F., Santofimia., M., & Lopez, J., 2011), and hybrid (Sanchez. & Saenz, 2008) (Sylvie Treuillet, 2010) systems. A comprehensive survey of the existing technologies for VIPs mobility was carried out in (Khan., S., Nazir., S., & Khan, H. U., 2021). Tools surveyed from this literature can further be divided in terms of their functionalities that mainly include VIPs' navigation, orientation, pathfinding, obstacle detection, object recognition, and scene interpretation. For these applications focusing on obstacle detection and object recognition in both indoor and outdoor environments, various hardware and software tools and devices have been utilised including digital cameras, electronic/traditional canes, radar, ultrasonic sensors, LiDAR, infrared, and thermal cameras (Khan., S., Nazir., S., & Khan, H. U., 2021). Recent works has been reviewed in this study specific to ODOMOVIP which propose various combinations of the aforementioned tools to assist VIP in mobility within their local surroundings.

A recent study (Parikh., Shah., & Vahora, 2018) presented a deep learning (DL) approach for object recognition for ODOMOVIP utilising a smart phone camera for capturing real-time images, which were streamed to a cloud-based server for processing to trained convolutional neural networks (CNN). The output of the proposed system provides a response to the user (via the internet) classifying the recognised objects by their names (11 objects in total). Similarly, (Shao., Han., Kohli., & Zhang, 2014) proposed stairs and person crosswalk detection using depth camera. The above methods used conventional image processing algorithms which can potentially be impacted by the real-time environmental dynamics (e.g., varying backgrounds and noisy images). To this end, more advanced generic pre-trained models such as YOLO (Redmon, Divvala, Girshick, & Farhadi, 2016), or Mask RCNN (He, Gkioxari, Dollar, & Girshick, 2017) could be considered with the ability to reliably and efficiently detect objects of multiple classes.

Work presented in (Bauer et al., 2020) utilises DL for object detection in an outdoor environment. Authors in this study proposed a hardware system based on wearable lightweight devices (e.g., camera, a smart watch used for triggering an alarm, coupled with a smartphone for connectivity) and developed a detection scheme for several objects including traffic lights, buses, cars, people, bicycles, and motorbikes with estimated depth map of the identified objects. Inference for the detected objects with respect to their distance and relative position is then forwarded to the user via spoken or haptic feedback. Another study (AtikurRahman & M. s., 2021) developed a tool for indoor and outdoor mobility of VIP using multiple laser sensors and a video camera for the detection and recognition of the objects of interest, respectively. Pre-trained models (MobileNet (Howard, 2017)) were used for object recognition where real-time data were streamed and stored on a remote server. While the object recognition outcomes from both studies indicate high accuracy, reported results for object identification were limited. Consequently, although indoor mobility assistance in the above studies is useful, consideration of outdoor objects specifically identified as potential hazards within the RNIB report (Wilson, 2015) can be a significant advantage for the above studies as part of the future work. Furthermore, detailed interpretation of a scene would be useful in describing the highest possible estimation of the depth, location, and distance of (a) specified objects within a smart city environment, and (b) challenging objects (e.g., e-Scooters, advertising boards) which have been placed dynamically and other static street objects for VIP.

In (Mattoccia, 2016), a DL based outdoor mobility assistance for VIP was proposed utilising an RGBD sensor and mobile device for semantic categorization of detected obstacles. Similarly, (Lin., B.-S., Lee., C.-C., & P-Y, C., 2017) proposed an outdoor mobility assistive technology for both online and offline scenarios. Pre-trained models including YOLO (Redmon, Divvala, Girshick, & Farhadi, 2016) were developed to detect objects which appear in front of the user walking trajectory. Despite the above advancements, conventional approaches for the estimation of distance (e.g., focal distance, pixel density, and height of camera device) are susceptible to false measurements due to the dynamics of real time environment which could lead to false alarm generation. In another study (Lin., Y., Wang., K., Yi., W., & Lian, S., 2019), a wearable system was proposed for the indoor and outdoor environment perception of VIP. The system used depth camera, processing unit, mobile based interface, and earphones to receive useful mobility information in real time. A DL model in the above study was developed for obstacle detection while using multiple secondary datasets for model training. After the detection of obstacles in the current frame/scene, the associated feedback is played back to user with varying sound volume proportional to the distance from the object (i.e., the higher the sound, the shorter the distance from the obstacle and vice versa). While the system indicated reasonable accuracy in varying lighting conditions, the outcome for VIP mobility is limited to only volume-based indications about the existence of identified obstacles rather than specified distance measures or identity of an obstacle.

Similar work presented in (Giarre., C. L., & al., e., 2019) uses a camera and inertial sensors embedded within a smart phone for indoor and outdoor navigation of VIP. Predefined paths were used with specific landmarks for real-time user navigation assistance. For tracking and localisation estimation, a Kalman filter was employed, producing reliable tracking performance. However, utilising specific markers such as corners or visual markers can potentially be affected in dynamic conditions such as occlusion and noisy backgrounds. Another wearable VIP mobility aid was proposed in (S. & P., 2015) comprising of 3D sensors mounted on glass frames (to capture depth map of the scene), a processing unit (for object detection), vibrotactile actuators, and bone conductive speakers (for delivering audio feedback to the user) generated by the system. Vibrations were generated for left, centre, and right-side locations of the detected object along with audio feedback of estimated distance from the obstacle. It should be noted that conventional image processing and computer vision algorithms might suffer from lower accuracy and reliability in the presence of real-time dynamics. Likewise, feedback generated could be comprehensive in addressing detailed scene information such as the name and size of the identified object, and geometry to be better perceived by a VIP.

The above literature indicates the use of sensor-based technologies with higher confidence in relation to VIP mobility assistance with a variety of possible approaches to support them in various aspects. However, several limitations can be faced including: a) lack of the detection and recognition of specific obstacles in outdoor environment that are identified by VIPs in RNIB report (Wilson, 2015); b) unavailability of the sufficient annotated data for corresponding objects; c) unreliable performance from the ObDtM specifically with real-time dynamics; d) limited perception and feedback for the identified scene and surroundings; and e) lack of the use of advanced tools and technologies specifically the synergetic behaviour of available technologies while considering the emergence of responsive IoT devices and smart city environments. Accordingly, a DTL approach was proposed in this study utilising multiple ObDtM models which were trained for a custom dataset compiled from various publicly available resources. The compiled dataset comprises of the objects of 8 different classes as identified within the RNIB report (Wilson, 2015) and considered as potential obstacles for the VIP during their outdoor mobility. Weights from two pre-trained models, namely YOLO (Redmon, Divvala, Girshick, & Farhadi, 2016) and Mask R-CNN (He, Gkioxari, Dollar, & Girshick, 2017) were used and tuned by using the above dataset. An interactive smart data driven mobility aid was the developed for the VIP (SOMAVIP) that utilises advanced tools and technologies enabling the interaction and mobility support for the ODOMOVIP by providing real-time comprehensive perception of the surroundings.

3. Methodology

A transfer learning approach was developed utilising well-established deep convolutional neural network structures of YOLOv5 and Mask R-CNNs based on a custom target dataset for real-time identification of 8 classes of objects (advertisement board, e-Scooter, wheel bin, trash bag, car, person, bollard, and bench). Development and validation of the proposed ObDtM model followed the workflow as described in Fig. 1 and explained in **Section 3.1**. Data for the objects of above classes were first acquired from various sources which was annotated and augmented to increase the size of target dataset size. Structures of pretrained DL models (i.e., body of the pretrained CNNs without the head and tail for identifying objects of new classes) were used for tuning and update of existing model hyperparameters to evaluate their performance on unseen instances of the above new domain. Model performance was then evaluated using various standard performance metrics as recommended by the original sources (He, Gkioxari, Dollar, & Girshick, 2017) (Jocher et al., 2020) which include: a) intersection over union (IOU - the intersection over the union for the ground truth and predicted bounding boxes), b) confidence score (CS) representing the probability that an anchor box contains a desired object, c) precision, d) recall, e) mean average precision and recall (mAP, mAR) for varying IoU and CS, f) F1 curve, and g) precision-recall curve. Detailed description with mathematical formulation of the selected performance metrics can be found elsewhere (Nick, 2018).

3.1. Data collection and preparation

There exist several commonly used datasets in relation to object detection such as COCO (Lin et al., 2014) (with multiple versions) and ImageNet (Russakovsky., O., & al., e., 2015) comprising of a large collection of annotated classes for 80 and 1000 objects, respectively. Although, these datasets possess a wide domain for object detection, they lack the objects of interest considered in the present study (i.e., objects identified by VIP in (Wilson, 2015)) such as wheel-bins, trash-

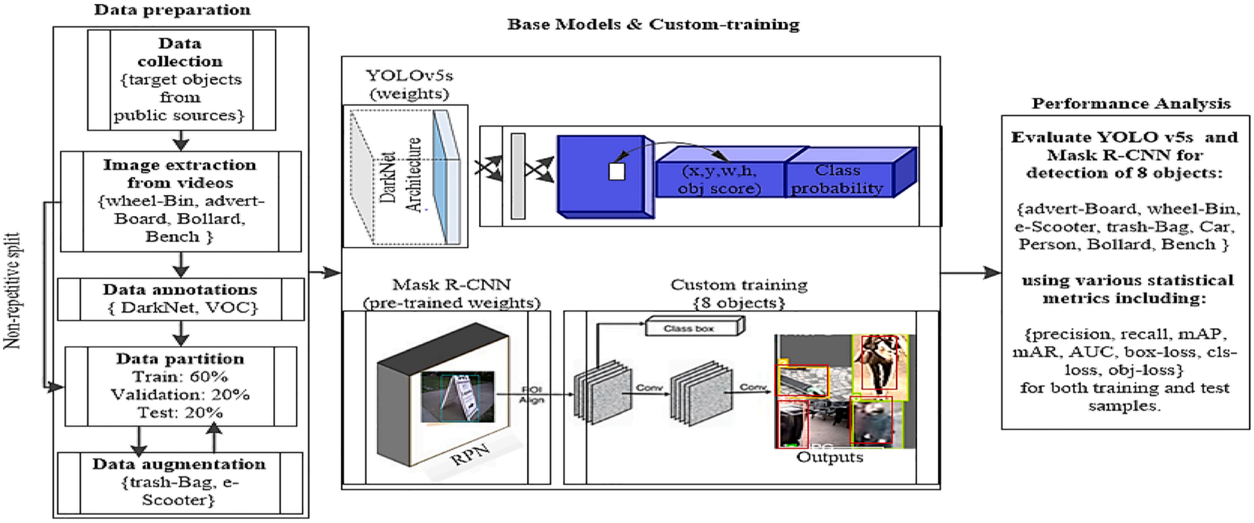
Data preparation
Data collection {target objects from public sources}
Image extraction from videos {wheel-Bin, advert-Board, Bollard, Bench }
Data annotations { DarkNet, VOC}
Data partition Train: 60% Validation: 20% Test: 20%
Data augmentation {trash-Bag, e-Scooter}
Non-repetitive split
Base Models & Custom-training
YOLOv5s (weights)
DarkNet Architecture
(x,y,w,h, obj score)
Class probability
Mask R-CNN (pre-trained weights)
RPN
ROI Align
Custom training {8 objects}
Class box
Conv
Conv
Outputs
Performance Analysis
Evaluate YOLO v5s and Mask R-CNN for detection of 8 objects:
{advert-Board, wheel-Bin, e-Scooter, trash-Bag, Car, Person, Bollard, Bench }
using various statistical metrics including:
{precision, recall, mAP, mAR, AUC, box-loss, cls-loss, obj-loss}
for both training and test samples.

Fig. 1. Workflow of custom YOLOv5s and Mask R-CNN model development for object detection identified by the VIP.

Bags, e-Scooters, advertising boards, and bollards *(i.e., roadside pillars; hereinafter termed 'pole')*. Additionally, sources with the annotations for these objects do not exist thus posing a challenge for implementing a transfer learning approach for the domain adopted in the present study. Accordingly, a publicly available 3D-scan dataset (without annotations) was acquired comprising of a variety of the required objects including benches, advertisement boards, poles, and wheel-bins (Choi, Zhou, Miller, & Koltun, 2016). 3D-scans in this dataset were captured via an R-GBD camera from varying orientations, distances, and angles producing detailed natural representations of data covering significantly high granularity as compared to typical augmented data generation schemes (e.g., zoom in/out, translation, rotation, and shear). 3D-scans of the required objects (bins, advert-Boards, poles, and benches) from this dataset were extracted and transformed to the corresponding image frames that were used for the custom training of the proposed ObDtM. For trash-Bags and e-Scooters, publicly available images (with NY-CC license) from Google image search engine were utilized comprising of a total of 520 {e-Scooter: 250, trash-Bag: 270} images without annotations. The annotated datasets for cars and persons were acquired from public sources (Krause., Stark., Deng., & Fei-Fei, 2013) and (DENG, Luo, Loy, & Tang, 2014), respectively. Images for the selected object classes were then annotated using a public annotation tool (DarkLabel: https://darkpgmr.tistory.com/16) producing the required data formats (bounding boxes, class label) to be used for ObDtM model. A summary of the final dataset used in the present study, including the selected object classes (identified by the VIP in the RNIB report (Wilson, 2015) and (Rnib, 2020)), sources, annotation statuses, and total number of compiled images is provided in Table 1.

Data annotations were manually generated via DarkLabel in the present study since they were not available from the original sources to train the custom ObDtM. Furthermore, relatively higher variations existed in the datasets with different image dimensions, backgrounds, and levels of resolution and orientation, reflecting the need for robust data management and processing techniques for the object detection in an ODOMOVIP. Data samples for e-Scooters and trash-Bags (i.e., 250 and 270 respectively) were significantly augmented and increased to a total of 2,594 images (e-Scooter: 1,152, trash-Bag: 1,442) after applying traditional data augmentation schemes.

It is noted that ill-defined or unordered data augmentation steps during model development process can cause model under or over training (i.e., model bias) by leaking critical information from test/validation to training sets when augmentation is performed inappropriately or prior to data splits. Accordingly, data annotation and partitioning (train/test/validation splits) were performed prior to the augmentation and transformation steps as illustrated in Fig. 1. Data partitions for training, validation, and test sets for both custom models (YOLOv5s and Mask R-CNN) were constructed with the ratios of 60%, 20%, and 20%, respectively as shown in Fig. 2. A randomly selected batch of annotated training samples for multiple object classes representing diverse properties (i.e., varying resolution, dimensions, translations, shear ranges, and backgrounds) and demonstrating ObtDtM model propensity for higher accuracy is presented in Fig. 3. The final dataset along with annotations is made publicly available (https://dx.https://doi.org/10.21227/sm6r-nb95) as described in supplementary information (S1).

3.2. *Proposed object detection (ObDtM) using transfer learning over custom data*

Object detection for a given task of ODOMOVIP with good accuracy

Table 1
Data distribution (ObDtM training and validation) and sources used in this study.

Object	Data source	Publicly Available	Annotated Originally	Total Instances (original)	No. of. Instances used for the ObDtM training & validation
Car	(Krause., Stark., Deng., & Fei-Fei, 2013)	Y	Yes	16,185 (196 types of cars)	2464
person (pedestrian)	PETA (DENG, Luo, Loy, & Tang, 2014)	Y	Yes	19,000	2544
e-Scooter	Google	Y	No	250	1152 (with augmentation)
trash-Bag	Google	Y	No	270	1442 (with augmentation)
advert-Board	(Choi, Zhou, Miller, & Koltun, 2016)	Y	No	112 3D-scans (videos)	2496
Bench	(Choi, Zhou, Miller, & Koltun, 2016)	Y	No	211 3D-scans (videos)	1984
Pole	(Choi, Zhou, Miller, & Koltun, 2016)	Y	No	70 3D-scans (videos)	2640
Bin	(Choi, Zhou, Miller, & Koltun, 2016)	Y	No	263 3D-scans (videos)	2432

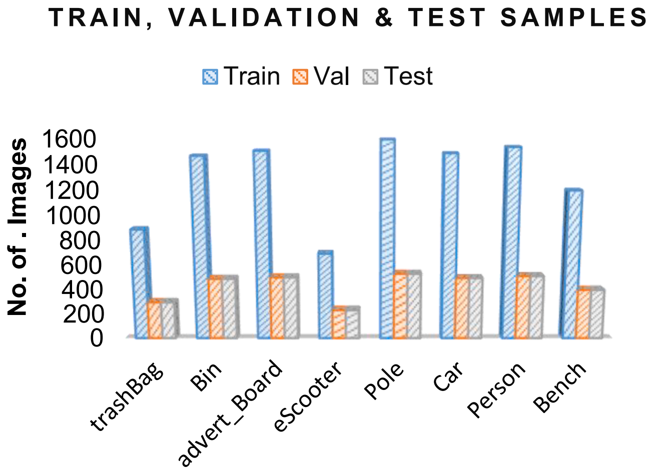

Fig. 2. Custom data distribution used for the training, validation, and testing of YOLOv5s and Mask R-CNN models.

can be accomplished via ***a)*** training a DL model from scratch, or ***b)*** using deep transfer learning (DTL). Constructing and training a custom new DL model requires excessive resources including large datasets, higher processing power (GPU for computer vision tasks), and thus longer training time. In DTL, pretrained CNN models constructed based on large existing datasets can be used to transfer the knowledge learned (model parameters) from one domain to another, even with lower data availability for the new domain, thus accelerating model learning processes (Pan & Yang, 2009). More specifically, DTL utilizes pre-trained models for application across contexts, thereby improving model generalization, and reducing the set of observations for the new domain and the training time required by conventional static machine learning approaches. To date, various pre-trained models have been proposed and made publicly available that can be utilised for the object detection, including noticeably YOLO (Redmon, Divvala, Girshick, & Farhadi, 2016), Faster-RCNN (Ren, He, Girshick, & Sum, 2015), Mask R-CNN (He, Gkioxari, Dollar, & Girshick, 2017) which extends Faster R-CNN, Single shot multi-box detector (W. & al., 2016), and Residual Network (ResNet) with various upgrades (from 20 to 101 convolutional layers) (He, Zhang, Ren, & Sun, 2015). Mask R-CNN is an instance segmentation DL model that aims to separate multiple objects in an image frame. In addition to bounding boxes and class names, it provides masks for a resulting image. Mask R-CNN first generates region proposals (RPN) for each object within the input image followed by its generation of the class level information and corresponding bounding box along with the pixel level mask for identified objects based on RPN from the first step. Mask R-CNN relies on its fundamental Feature Pyramids Network (FPN) (Lin.,

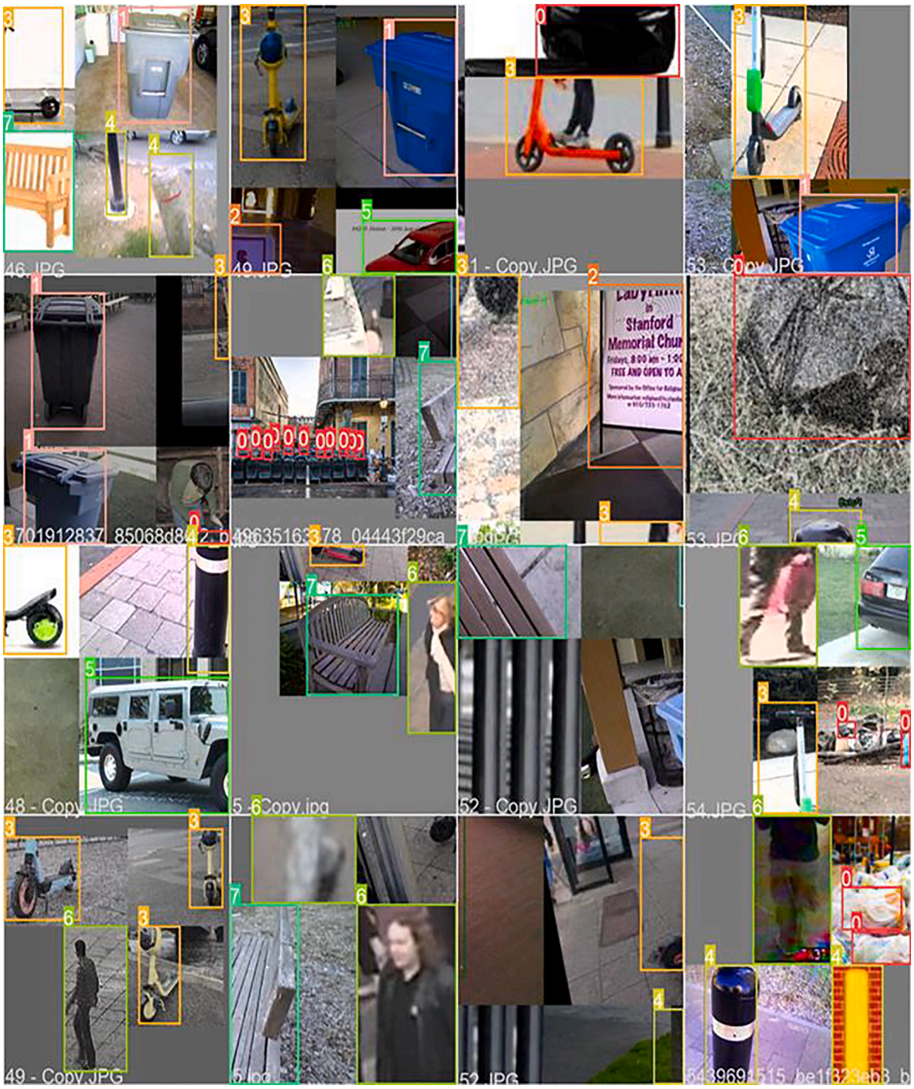

Fig. 3. A randomly selected batch of training samples from custom annotated data comprising objects presenting diverse perspectives and geometric properties.

T.-Y., Dollar., P., Girshick., R., He., K., Hariharan., B., & Belongie, S., 2017) approach that is used for the object detection in images of varying scales. Varying scale in FPN is beneficial over single CNN which has demonstrated to maintains robust semantic features. Mask R-CNN was originally trained over COCO dataset (Lin et al., 2014) comprising 80 object classes, with 2.5 million labelled instances in 328,000 images and has extensively been adopted in various areas focusing on object detection and scene recognition. Detailed description of Mask R-CNN implementation, configuration, and mathematical formulation can be found in the original work (He, Gkioxari, Dollar, & Girshick, 2017).

The family of R-CNN models has been shown to perform with lower efficiency for real-time object detection tasks when compared with YOLO (Redmon, Divvala, Girshick, & Farhadi, 2016) due to their multi-stage processing mechanism. In contrast, YOLO contains a single CNN to detect the position and class of a desired object due to the parallel embedding of the classification operation at each layer of the network convolution operation at each layer. YOLO therefore considers object detection as a single regression task directly from the input image to the predicted objects' locations (i.e., bounding boxes) with associated class probabilities. Similar to other modeling approaches for object detection and recognition tasks, YOLO has been upgraded with several sequential improvements in the form of versions and made publicly available as open-source models (Redmon & Farhadi, 2018). YOLOv5, the latest version (Jocher et al., 2020), was trained based on COCO dataset (Lin et al., 2014) and is available with three variants of small, medium, and large network structures. Although reported with performance degradation issues for detection/recognition of objects of smaller sizes or the objects located in closed proximities with the image (Cao, Liao, Song, Chen, & Li, 2021), YOLO model family has shown to be among the most efficient detection models for real-time scene interpretation. Additionally, recent literature has reported upgraded YOLO models (DarkNet-53 where 53 is the number of convolutional layers) to be faster than the family of ResNet models (Srivastava et al., 2021). Furthermore, YOLOv5s models were compared with Mask R-CNN with respect to several statistical measures where Mask R-CNN models are an extension of the Faster R-CNNs which utilize ResNet (ResNet-50/101) as their backbone structure. Irrespective of the better YOLOv5s model performance presented in this study, the proposed YOLOv5s can also be easily extended to incorporate residual layers (ResNet methodology) for further performance improvement by: (i) increasing batch input size, and (ii) reducing the spatial resolution of image sizes and maintenance of gradient information (computed within the network), thus effectively handling the flow of gradient and avoiding vanishing/exploding gradient issue (Tan, Huangfu, Wu, & Chen, 2021).

In the present study for object detection tasks targeted for VIP and a proof of concept for SOMAVIP, the ObDtM was developed using pre-trained mask R-CNN and YOLOv5s models and tuned (model hyper-parameter update) based on a custom dataset comprising of eight objects presented in Table 1. As shown in Fig. 1, weights of these pre-trained model were used which were initially learned based on large datasets (COCO and ImageNet) to utilise prior knowledge. For training and testing of custom object detection, a computational machine with GPU (AMD Ryzen Threadripper 2990WX 32-Core Processor 3.00 GHz, 128 GB RAM) was used. Further details of the configurations of these models along with the proposed ObDtM are presented in supplementary information (S2). Model performance was evaluated for varying training and validation epochs with a maximum allowed number of 120 epochs.

4. Results and discussions

Performance of the custom YOLOv5 model for the selected domain of 8 object classes with respect to the final epoch over validation data and purely unseen test dataset is presented in Table 2. It can be seen that the developed DTL model achieved higher performance for training data (mAP@0.5: 97%) with the exception of moderate performance of 87% mAP@0.5 for trash-Bag. This may be partly due to: a) a potential YOLO tendency for the *person, car,* and *bench* classes given that a larger proportion of the samples for these classes were available in COCO dataset, or b) treatment of the trash-Bag as a relative minority class due to the limited availability of the data for this class, therefore, requiring more training of the custom model over trash-Bag instances.

Performance of the ObDtM model with respect to the selected statistical measures, particularly the mAP@0.5 and mAP@0.5-0.95 (i.e., IoU = 0.5–0.95), based on training and validation sets demonstrated continuous improvement for increasing number of epochs (120 epochs in total) reaching a plateau approximately after 60 epochs (Appendix in S1). It is however noted that a possible bias (section 3.1) may have caused higher relative model validation performance of the model for the objects {bench, sign board, bollard, wheel bin} due to the likeliness of higher number of similar instances of these objects in the custom dataset that were extracted from the 3D-scan videos for both training and validation sets.

This can be observed in class level mAP generated during model training phase as given in Fig. 4. Nearly all the object classes were classified with a mAP > 0.95 with an exception of trash-Bag that was 87% which can potentially be due to a lower number of samples for trash-Bags (minority class). Despite the lower relative performance indicator for the trash-Bag class, an average mAP@0.5 of 97% for the overall dataset was accomplished demonstrating good model performance. Higher overall model performance is evident of the robustness of YOLOv5s model due to the fact that the weights initially trained on a large COCO dataset of a multitude of object classes with similar features (Lin et al., 2014) were adopted for relatively lower domain area of 8 object classes, thus indicating successful transfer of the learned parameters with adequate size of the target dataset. For example, weights for the object classes {car, person, bench} were already trained within the original YOLO model, thus demonstrating 99% mAP over unseen instances for the new samples of the objects of these classes.

Performance of the developed custom YOLOv5s final model (i.e., trained and evaluated based on training and validation sets, respectively) was evaluated for the test set where individual images as well as the frames extracted from objects' videos, kept unseen in model training and validation phases, were used (See Fig. 1). An average of 95% and 96% precision and recall were reported for all object classes in the test

Table 2
Performance of custom-trained yolov5s for validation (during Training) and purely unseen test data for 8 objects.

Object	Precision		Recall		mAP(0.5)		mAP(0.5–0.95)	
	Train	Test	Train	Test	Train	Test	Train	Test
car	0.98	0.95	0.99	0.99	0.99	0.99	0.87	0.84
person	0.99	0.98	0.99	0.95	0.99	0.99	0.99	0.99
e-Scooter	0.99	0.93	0.97	0.89	0.98	**0.93**	0.73	0.65
trash-Bag	0.75	0.89	0.87	0.91	0.88	**0.94**	0.64	0.71
advert-Board	0.99	0.98	1	0.99	0.99	0.99	0.98	0.99
bench	0.98	0.99	1	0.95	0.99	0.98	0.97	0.97
pole	0.98	0.93	0.99	0.97	0.98	0.95	0.97	0.94
bin	0.98	0.99	1	0.98	0.99	0.99	0.99	0.99
Overall	**0.96**	**0.95**	**0.97**	**0.96**	**0.97**	**0.97**	**0.89**	**0.89**

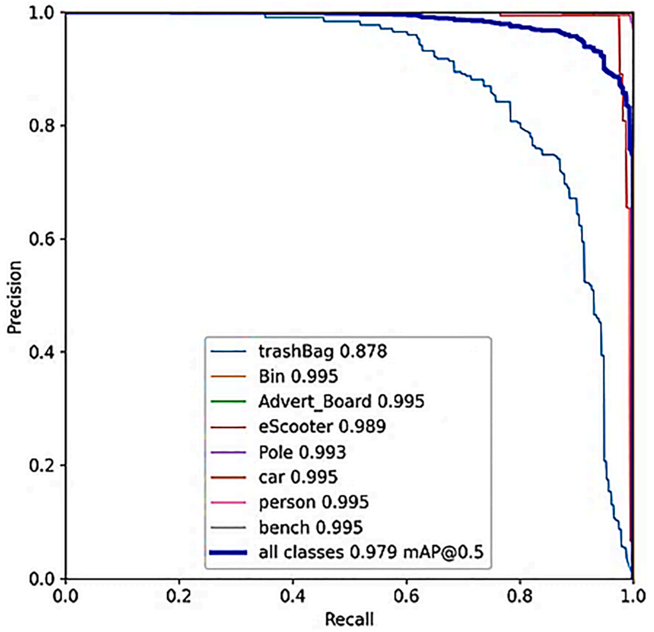

Fig. 4. Performance of the custom YOLOv5s model with respect to precision-recall curve for training dataset along with the retrieved mAP@0.5 (i.e., IoU = 0.5).

dataset, with the lowest of 89% and 93% precision and recall observed for trash-Bag and e-Scooter, as given in Table 2. This is congruent to the outcomes observed for the training dataset (in Table 2) which also indicated a somewhat lower mAP for these two classes. The mAP@0.5 indicates overall 97% mAP for all classes while<95% for e-Scooter (93%) and trash-Bag (94%) and over 95% for remainder of the individual classes. Furthermore, mAP@0.5–0.95 also showed 89% mAP, with e-Scooters and trash-Bag as least performers (65% and 71% respectively). Overall, the alignment of these outcomes with the training performance (Table 2) clearly indicated model generalisation and reliability when tested over unseen samples.

F1 measures for all objects with varying confidence score is presented in Fig. 5**(a)** indicating an optimal F1 score of 96% with confidence of 32% which remains stable until 80% confidence, indicating YOLOv5s model robustness with good accuracy. For objects of e-Scooter and trash-Bag classes, an improvement in the prediction performance can certainly be achieved by acquiring more training samples. Fig. 5**(b)** shows the trade-off between the precision and recall metrics indicating the best compromise overall for all objects listed in Table 1 producing 97.5% mAP@0.5. These outcomes clearly indicate the reliability and generalisation of the custom trained ObDtM and hence its suitability to be embedded in the proposed SOMAVIP.

The second set of experiments was based on the adoption and tuning of weights from pretrained Mask R-CNN model (originally trained for COCO dataset [50]) for custom dataset acquired in this study (Table 1). The predefined configuration of the Mask R-CNN model was used as recommended in the original study (He, Gkioxari, Dollar, & Girshick, 2017) while training the fully connected layer over custom data listed in Table 1. Training, validation, and test sets of the same configuration (60%, 20%, 20% splits) as for the YOLOv5s were kept (Fig. 1). The model was retrained for 100 epochs for the training set (Table 1). Performance of the trained Mask R-CNN model reached a plateau at 70 epochs without a noticeable improvement for training and validation sets (Appendix in S1).

Statistical outcomes for the final epoch (epoch 100) for the Mark R-CNN model, recorded for both the validation and test sets are presented in Table 3. Overall mAP@0.5 for training set was of 92%, with the degradation of 5% when compared to 97% mAP@0.5 of YOLOv5s model. More specifically, mAP@0.5 was dropped for the objects of classes {bin, e-Scooter, pole, bench} whereas no noticeable change was observed for {trash-Bag, advert-Board, car, person}. Additionally, a significant reduction in the mAP(@0.75 to 85% was noticed as compared to YOLO based object detection which was of 90% even when evaluated at a higher threshold (@IOU:0.95). Performance of the final Mask R-CNN model (i.e., trained and validated/evaluated for training and validation sets, respectively) was then evaluated for unseen test set as described in Figs. 1 and 2.

Results for the test data objects are also shown in Table 3 indicating the overall mAP@0.5 of 93% with the minimum mAPs of 80% and 81% recorded for the e-Scooter and trash-Bag, respectively, whereas a perfect score of 100% for {cars, person, bench} classes. Overall, the mAP@0.5 was reduced from 97% (YOLO based model) to 93% for Mask R-CNN which may be significant when considering the reliability required in real time applications such as within the proposed SOMAVIP. Furthermore, noticeable class-level performance differences can be seen between the two selected modelling approaches, particularly for the trash-Bag (reduced from 94% in YOLOv5s to 81% in Mask R-CNN) and e-Scooters (reduced from 93% in YOLO to 80% in Mask R-CNN). This is evident of the reliability and generalisation of YOLOv5s custom-trained

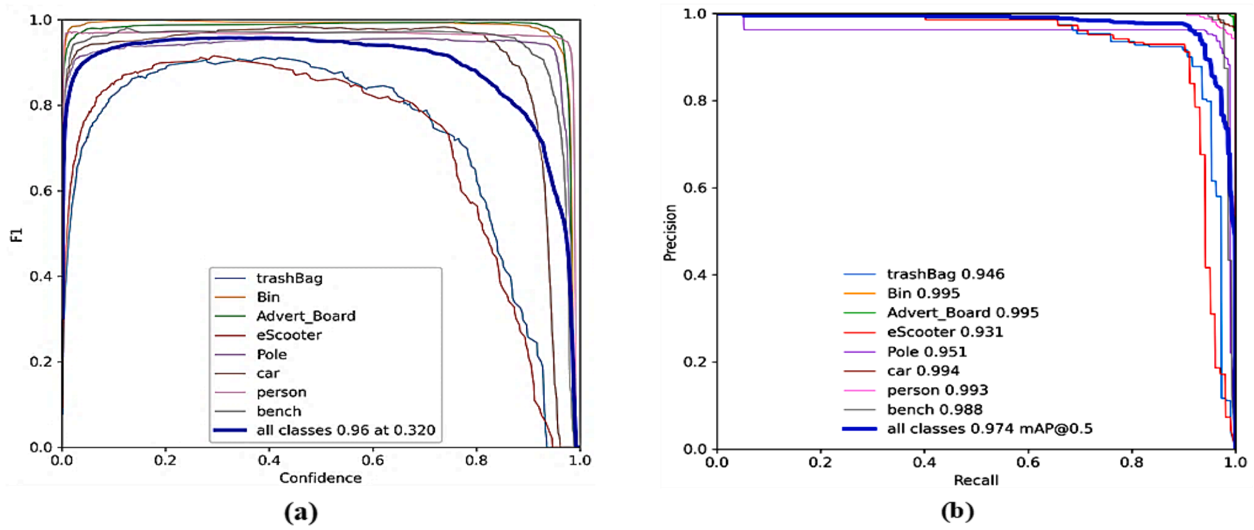

Fig. 5. **(a)** F-1 Curve for YOLOv5s for the test dataset showing F1 measure profile with respect to the varying confidence score for each class in unseen test dataset, **(b)** YOLOv5s precision and recall curves for mAP (0.5) for unseen test data samples.

Table 3
Performance of custom trained Mask R-CNN over the unseen test set for 8 classes (objects).

Objects	mAP(IoU = 0.5)		mAP(IoU = 0.75)		mAR(IoU = 0.5)		mAR(IoU = 0.75)	
	Train	Test	Train	Test	Train	Test	Train	Test
car	0.99	1	0.95	0.93	0.99	1	0.98	0.97
person	0.98	0.99	0.98	0.99	0.99	1	0.99	1
e-Scooter	0.85	0.80	0.58	0.56	0.90	0.87	0.70	0.64
trash-Bag	0.85	0.81	0.56	0.69	0.88	0.87	0.63	0.71
advert-Board	0.98	0.98	0.96	0.89	0.97	0.98	0.96	0.93
bench	0.78	1	0.75	0.99	0.83	1	0.77	1
pole	0.88	0.95	0.85	0.91	0.91	0.97	0.86	0.92
bin	0.92	0.89	0.91	0.85	0.94	0.89	0.91	0.87
Overall	**0.92**	**0.93**	**0.86**	**0.87**	**0.93**	**0.95**	**0.85**	**0.88**

model for the present study. Finally, a slight performance degradation from 85% mAP for training set to 80% for test set (Table 3) for {trash-Bag, e-Scooter} was observed indicating the need for acquiring additional domain knowledge for objects of these classes.

Finally, YOLO based object detection model is advantageous over the family of Mask R-CNN models (Redmon, Divvala, Girshick, & Farhadi, 2016), with respect to the efficiency due to its architectural simplicity which can particularly be useful for applications of real time mobility aid in a smart city environment. YOLOv5s and Mask R-CNN modes were also evaluated with respect to the processing times for test data where, on average, YOLO based custom-trained model required 2.8 ms to preprocess and perform inference on an image of the dimension (640 × 640 × 3) as compared to 0.002 s per image frame by Mask-R CNN. The difference is substantial for the problem in hand (i.e., ODOMOVIP) with high throughput requirement for real-time scene perception while processing live video streams.

In summary, the selected custom-trained models showed merit in terms of generalisation for the targeted object detection when evaluated for unseen instances of the test sets with varying noise, backgrounds, resolution, image quality, and orientation. On average, YOLO based DTL outperformed Mask R-CNN with respect to the prediction accuracy (based on the selected statistical measures) and the required processing time. YOLOv5s for the custom dataset demonstrated a robust modeling approach with wider applicability domain where detection for several frames per second may be necessary with the feedback required in real time for the ODOMOVIP.

Although several object detection techniques have been proposed in the literature in relation to ODOMOVIP, significant challenges still remain with respect to reliable and accurate real time object detection, and importantly the lack of availability of the specific domain knowledge (e.g., objects of 8 selected classes) as identified by the VIP in RNIB report and other sources (Wilson, 2015) (Rnib, 2020). A comparative analysis was also conducted in the present study, as shown in Table 4, summarizing the differences between the proposed ObDtM and recently introduced VIPs' mobility aids for outdoor and indoor environments, corresponding methodologies, detected objects, sensors used for data capture, and the reported outcomes. Although several studies in Table 4 emphasized on the use of computer vision and DL algorithms reporting mAP and classification accuracy up to 77% and above 90%, respectively for different object classes, majority of these studies are based on conventional object recognition approaches in contrast with object detection methods that are most appropriate for the proposed study of ODOMOVIP. Additionally, image segmentation using the reported conventional methods may potentially suffer from reliable and robust outcomes in outdoor dynamic environment such as occlusions, poor lighting, and the effects of image distortion from weather conditions. Existing works may therefore lack the capability for objects detection with higher accuracy that are identified by the VIP (Wilson, 2015) (Rnib, 2020). Furthermore, the present study is advantageous with the provision of annotated datasets and custom trained object detection models for the advancement of studied area of research and development. Finally, existing tools are unable to provide comprehensive feedback to VIP that could entail a detailed scene information (such as name and size of identified object, distance from VIP, orientation, current state, and geometry) for smart perception to VIP.

Finally, despite the robust performance of proposed custom-train ObDtM for the given classes of obstacles, there are some limitations in this study which can be considered in ongoing future works. For instance, image dataset can be enlarged with additional types of obstacles particularly, potholes, puddles, trash-Bags, e-Scooters, and scaffolding that are identified in the literature, as potential obstacles for the ODOMOVIP. The custom-trained ObDtM can be then fine-tuned over the annotated instances of additional classes. Likewise, it can be noticed that the overall performance of ObDtM is comparatively lower in case of trash-Bags and e-Scooters. This is mainly because both of these classes comprise less number of image instances (i.e., minority classes). This clearly indicates the need of additional dataset for such classes that will further improve the reliability and generalisation of our custom-trained ObDtM.

5. Proposed smart outdoor mobility framework for VIP (SOMAVIP)

The components of a proposed smart SOMAVIP were outlined emphasizing the need for robust and efficient real-time object detection techniques to develop a comprehensive synergetic assistive framework as a critical tool to enrich VIPs perception of their surroundings. Accordingly, accomplishing in the proposed study with custom-trained YOLOv5s and Mark R-CNN (Tables 2-3) were pinpointed for specified objects, demonstrating a robust and interpretable approach of ObDtM when evaluated for a dataset of unseen instances with dynamic conditions. In this regard, an end-to-end framework is proposed in Fig. 6 for real-time scene interpretation and surrounding perception for ODOMOVIP while considering the future data driven smart IoT and responsive environments.

As described in **Section 4**, existing assistive technologies for ODOMOVIP are in need of a significant technical and hardware advancement to cooperate with emerging future responsive devices specifically those concerning the VIPs mobility in a synergetic smart city environment (Khan & Kuru, 2021). Domain knowledge and decision support for the real time scene interpretation provided by existing tools (e.g., (Bauer et al., 2020), (Lin., B.-S., Lee., C.-C., & P-Y, C., 2017), (Giarre et al., 2019)) require further enhancement by incorporation of synergetic behaviour and information from standalone tools to build next generation ODOMOVIP as a harmonious part of future smart city environment. Accordingly, a detailed framework focusing on the critical building blocks of the proposed SOMAVIP are illustrated in Fig. 6, comprising of a set of mobile device and headphones as the required hardware components. General off-the-shelf components of the proposed SOMAVIP are described in the following sub-sections.

5.1. *Static text (labels)*

A collection of lightweight static text (static labels such as 'there is')

Table 4
Comparison between state-of-the-art object detection frameworks for VIP outdoor mobility and proposed SOMAVIP comprising the custom-build ObDtM.

Study	Main objective and approach	Sensors used	Type	Objects detected	Objects from RNIB list	Outcomes and limitations
(Parikh., Shah., & Vahora, 2018)	DL and image processing for image classification. No specified feedback about surrounding is presented.	2D images from smart camera; cloud-based remote services for data processing.	Outdoor	bench, bicycle, car, dog, motorbikes, person, pole, stair, traffic signals, trees, walls.	4 objects from RNIB list (pole, person, bench, car)	Varying classification accuracy. 81%(min) to 99%(max) for different objects. Object detection needs improvement. Dataset and annotations not provided.
(AtikurRahman & M. s., 2021)	SSD-based object recognition. Feedback about distance (near, intermediate, far) is provided to user	Laser sensor and camera device; accelerometer, cloud based remote services	Indoor & Outdoor	Currency notes (8 types), 5 objects (person, stairs, chair, table, washroom)	1 object from RNIB list (person)	Indoor and outdoor object recognition accuracy 98.11% and 98.7% respectively. The mAP measure not reported.
(Bauer., Z., & al., e., 2020)	YOLOv2 for object detection. Information about positions of potential obstacles is fed to user as haptic or spoken feedback.	Images from camera; remote server, smart watch	Outdoor	Bicycle, bus, car, motorbike, person, traffic light, traffic sign	2 objects from RNIB list (person, car/bus)	87.9% mean accuracy in obstacle presence detection. 74% mAP for 7 objects that needs improvements for real application. Dataset and annotations not provided.
(Shao., Han., Kohli., & Zhang, 2014)	Image processing-based detection and recognition of stairs and pedestrian crosswalks	RGBD camera	Outdoor	Stair, pedestrian crosswalk	0	93% detection accuracy, 95.8% classification accuracy
(Mattoccia, 2016)	DL and image processing based semantic segmentation of obstacles. Haptic and audio feedback to user	RGBD camera, mobile device, processing unit	Outdoor	Bench, car, poles, person, steps, trash-can, trees, walls	4 objects from RNIB list (pole, person, bench, car)	98% obstacle detection accuracy, 72% classification accuracy that needs improvement. Also, object detection would be more appropriate as compared to object classification. Dataset and annotations not provided.
(Lin., B.-S., Lee., C.-C., & P-Y, C., 2017)	DL for object detection in front of user when walking while providing object identity and distance as feedback to user. Propose online & offline mobility assistance.	Mobile device, camera, remote server for data processing	Outdoor	Bike, bus, car, motorbike, person, pier, potted plant	2 objects from RNIB list (person, car/bus)	mAP 20% to 60% when tested over unseen data which require significant improvements. Dataset and annotations not provided.
(Lin., Y., Wang., K., Yi., W., & Lian, S., 2019)	DL for object detection. Feedback about semantic map via varying sound's volume representing distance from object.	RGBD camera, processing unit, mobile interface, earphone	Indoor & outdoor	Multiple low-laying indoor (wall, floor, cabinet, bed and chair, etc.) and outdoor (road, sidewalk, person, car and building) objects	3 objects from RNIB list (person, car/bus, bench)	96%-99% accuracy for collision instruction classification. Object detection results (such as mAP) not reported that is the major limitation. Dataset and annotations not provided.
(Giarre et al., 2019)	Uses predefined paths with specific landmarks to be detected using image processing. Kalman filter is used for the user tracking.	Camera, inertial sensors, smart phone	Indoor & outdoor	Special landmarks such as corners, visual markers	0	User heading error: 5-degree, Object detection results not reported. Dataset and annotations not provided.
(S. & P., 2015)	Conventional image processing algorithm for obstacle detection. Generate vibrations for objects and audio for distance from object.	3D sensor, glass frame, processing device, actuators, speakers.	Indoor & outdoor	No object detection is performed. Obstacle within specified region of interest	0	Statistical results not reported. Dataset and annotations not provided.
ObDtM (Our model)	**Custom-trained YOLOv5s and Mask R-CNN to recognise and localise objects**	**2D images (camera)**	**Outdoor**	**8 Objects reported by VIP (see Table 1)**	**8 objects from RNIB list including e-Scooters (see Table 1)**	**97% mAP (for overall 8 objects). Dataset along-with annotations are provided.**
Proposed SOMAVIP Framework	**Comprehensive perception and commentary about surroundings within certain range (both offline & online). Advanced DL-based object and obstacle detection identified by the VIP as potential hazard for ODOMOVIP.**	**Mobile device, LiDAR, cloud services, smart, IoT, static information, headphones,**	**Outdoor**	**Various objects and obstacles tar are potential hazard form ODOMOVIP**	**Obstacle detection & recognition, object detection**	**Similar to our ObDtM (97% mAP), because SOMAVIP components are established e.g., smart devices, IoT, cloud services, LiDAR, navigation**

is required to be embedded within the outcomes generated from cloud-based object interpretation and obstacles' geometric map construction services. For instance, cloud-based services may infer the class, orientation, dimensions (height, width, and depth), and distance of the objects from the perceived location, which should be integrated with 'static' labels to construct a complete sentence (commentary) for the surroundings environment of interest. Additionally, static labels for fixed obstacles such as street premises (e.g., phone booth, street benches) may be useful for scene interpretation with discrete level information for the ODOMOVIP.

5.2. Scene interpretation

Scene interpretation may be accomplished within the local mobile device and will require the integration of useful information from all ODMOVIP components as shown in Fig. 6. Interpretations retrieved

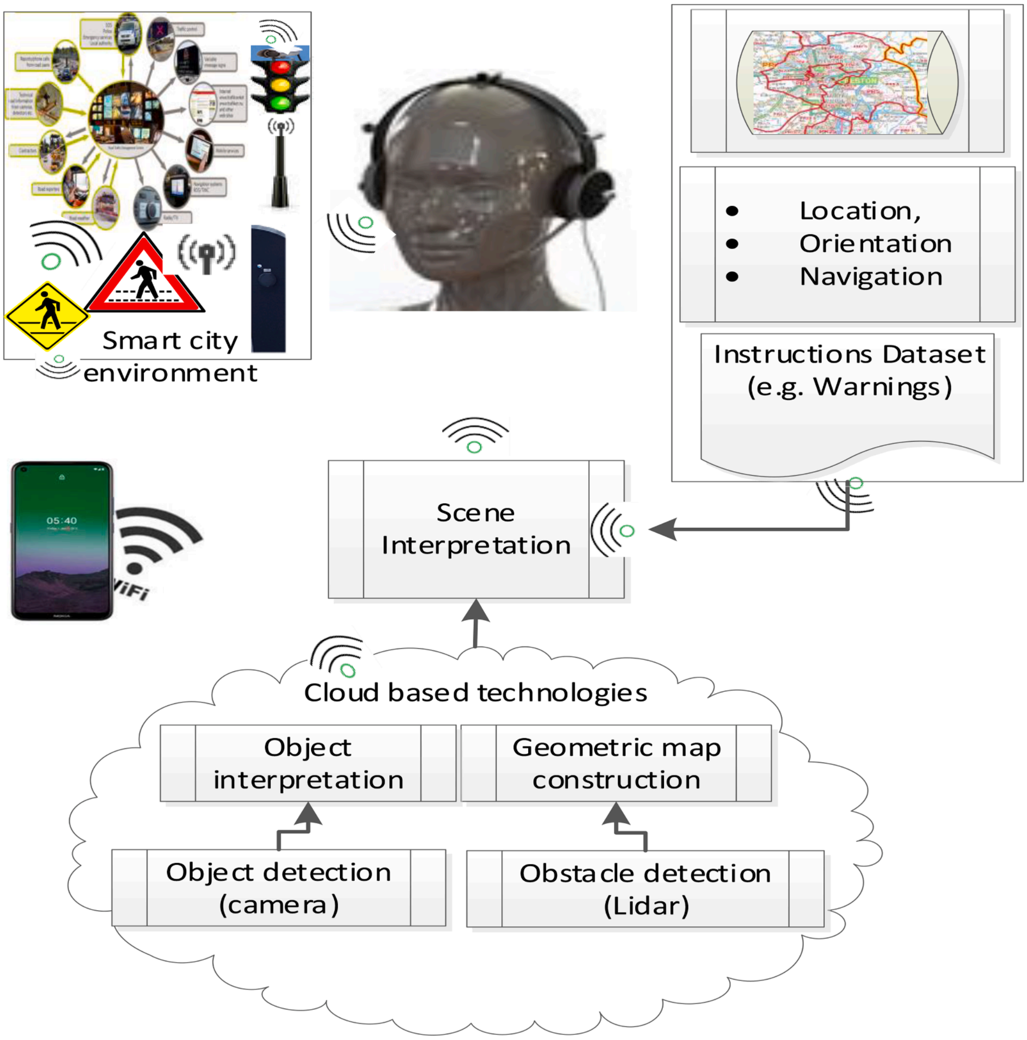

Fig. 6. Building blocks of the proposed smart SOMAVIP framework for perception and interpretation of surroundings.

from cloud services (i.e., object and obstacle identification and geometric map), smart phone (e.g., location, orientation, navigation), static text and information (e.g., from local-authority databases, (through Google maps), as well as real-time IoT devices (specifically in smart city context), will be merged together to generate a real-time commentary (perception) for the surroundings within the given constraints (e.g., only front view, within certain distance, and specific sampling frequency) to be perceived by the VIP during their ODM. The information generated from the above will then be transmitted to VIP via wireless headphone device with an embedded third-party speech synthesis support (e.g., Google services in smart phones). Furthermore, by utilising third-party services (e.g., Google translation), commentary can be instantly transcribed and translated into multiple languages allowing usability across the globe for a wider linguistic audience.

5.3. Smart mobile device

Smart mobile devices can be useful to produce reliable orientation, location, and navigation information of the end user, as presented in (Shao., Han., Kohli., & Zhang, 2014) and (S. & P., 2015). In the proposed SOMAVIP, a smart mobile device will be used for (i) live video streaming at a specified frequency (to remote server) via wireless network, (ii) navigation services (in local neighbourhood), and (iii) communication of the device with the end user (via speech recognition and synthesis) as well as smart responsive devices and IoT (for a smart city environment). In case of a future smart city IoT infrastructure, real-time information captured through IoT devices may also be utilised for further processing and smart perception of the surroundings in addition to the scene interpretation for the VIP.

5.4. Cloud based services

As proposed in (Lin., B.-S., Lee., C.-C., & P-Y, C., 2017), information captured through sensor devices can be processed either on a local device or through remote services. With the generation of voluminous data through the use of data driven technologies, IoT, and smart city environments, storage and efficient processing of the domain knowledge with higher response time and availability via cloud based services can be a significant undertaking for decision support in a variety of applications. Recent advancements in cloud computing infrastructures have been proven useful for handling big data generated from digital environments and have been utilised in diverse application domains (Khan & Kuru, 2021) (Yang., C., Huang., Q., Li., Z., Liu., K., & Hu., F., 2017). Accordingly, the proposed SOMAVIP will rely either on an offline service (i.e., high specification mobile device) or cloud-based platforms for real-time processing of the data generated via multiple sensors and other IoT sources. As shown in Fig. 6, cloud-based infrastructure will mainly host the following two main components:

5.4.1. Object detection model (ObDtM)

A pre-trained ObDtM (e.g., YOLOv5s in **Section 2**) hosted over the cloud with high availability and efficiency will provide instantaneous and automated object classification and localisation of objects identified within a specified field of view (e.g., front view as used in (Lin., B.-S., Lee., C.-C., & P-Y, C., 2017)). The outcomes from ObDtM as a sub-major SOMAVIP components will then be fed to the interpretation modules to transcribe and translate the results into desired semantic/textual context (e.g., object name, location, size). The custom built ObDtM can robustly identify and localise objects of multiple classes in real-time video stream as the core component of the proposed SOMAVIP and therefore a critical element of the cloud services. Furthermore, in case of a newly identified object that can a potential hazard for the ODOMOVIP, the ObDtM can be fine-tuned offline over additional image instances (for new class) to update the online ObDtM hosted over cloud.

5.4.2. Obstacle detection model

Similar to ObDtM, the obstacle detection component will comprise a pre-trained DL model to process the LiDAR data stream and generate the geometric map and semantics of the obstacles and identified objects (by ObDtM). There have been several recent research advances in relation to scene interpretation. For instance, a deep stereo geometric network is proposed in (Chen, Liu, Shen, & Jia, 2020) which detects 3D objects on a differentiable volumetric representations. Alternatively, (Rist, Emmerichs, Enzweiler, & Gavrila, 2022) proposed semantic scene generation using DL and LiDAR data which shows comparatively reliable outcomes. Likewise, (Zhao, Pang, & Zhang, 2018) introduced objects' shape extraction from LiDAR scans of outdoor scene. These works clearly indicate the usefulness of LiDAR based scans particularly for the automatic generation of objects' semantics in the form of geometric maps. Such outcomes can be used in the proposed SOMAVIP as one of the major component along with the ObDtM to interpret the geometric properties of the detected obstacle (e.g., depth, distance, shape etc.). The combined information generated from the ObDtM and LiDAR based semantic representations will be useful for the ODMOVIP by producing detailed interpretations of the surrounding environment and particularly, the obstacles and objects that are the potential hazards in their mobility.

Sequential procedure of the proposed SOMAVIP framework is presented in the algorithm (Algorithm 1) which is expected to generate synthesised outcomes comprising comprehensive information and perception about the current temporal and spatial state of the surroundings.

Algorithm 1. End to end procedure in proposed SOMAVIP

```
Let $C$ be the cloud-based pre-trained DL models for the object detection ($D_1$) and
  obstacle detection ($D_2$)
Let $T$ be a set of static text/information to be used for the commentary
Let $V$ to be the set of validation parameters for input sensors {e.g., sample rate, image
  resolution, view ranges etc.}
Let $O$ be the output perception & interpretation about the detected objects, obstacles,
  and surroundings
Let $L$ be the list of newly identified objects that can be hazardous for the ODOMOVIP

Procedure:
IF $L$ is NOT empty:

- Collect additional image samples for the object/s contained in the $L$
- Perform the Offline fine-tuning of the $D_1$ (i.e., ObDtM)
- Replace the cloud-based $D_1$ with the updated $D_1$
Else:
    For each video frame/s $F$ and sensor input $S$ for a given time interval $t$:

- Stream $F$ and $S$ to $C$
- Validate $F$ and $S$ using $V$
- Let $L_1 = [\ ]$ and $L_2 = [\ ]$ are empty lists to store the identified objects and obstacles
  respectively
- IF valid input from sensors:
  - Use the $D_1$ and store the identified objects in list $L_1$
  - Use the $D_2$ and store the identified obstacles in $L_2$
  - For each detected object and obstacle in $L_1$ and $L_2$
    - Measure the pre-defined features {e.g., location, colour, length}
    - Construct the geometric map {e.g., shape, depth, width, height, distance}
```

(continued on next column)

(continued)

Algorithm 1. End to end procedure in proposed SOMAVIP

```
    - Store the generated interpretation in a vector $I$
    - Return $I$ to mobile device via wireless connection
- End loop
      End IF

- Check (if there exists) dynamic state information {e.g., smart IoT devices} at time t
  and validation set $V$
- Get information (if there exists) about local outdoor premises {e.g., local authority
  database for roadside premises}for the current location $L$ and within certain
  distance in $V$For each item in $I$
  o Embed information from $I$ into corresponding entry in $T$
  o Embed the dynamic state information into $T$
      End loop

- Generate the perception & textual commentary $O$ for current time $t$ using $T$, $I$, and
  IoT & dynamic information
- Transform the textual perception into spoken commentary via 3rd party speech
  synthesis
- Feed the audio commentary to user via wireless headphones
End loop
```

Some examples of expected outcomes are briefly described below:

"Warning! at < 10 > feet distance, a < red > car on your < top-left > is < parked > with estimated length of < 4 > meters".

"Warning! an < e-Scooter > is < laid > on < 30 > degrees on your < top-right > side in < 2.5 > meters range and there is a < green> <wheel-bin > in < front > of you in < 1 > meter range with estimated height and width of < 1.2 > meters and < 0.7 > meters respectively".

The RNIB survey has reported over 90% collisions of VIP with obstacles over a short period of three months. There are over 50% identified outdoor dynamics (e.g., new developments in local area) which pose significant challenges for ODM (Wilson, 2015). Additionally, over 33% of the reported injuries of VIP may directly or indirectly increase the associated costs including medications, injury claims, and other related treatments. Furthermore, reduced ODM in VIP may affect their mental health, exacerbate the risk of dementia, and raise the need for additional social care, affecting the overall quality of life. Accordingly, a fully autonomous SOMAVIP framework will be of significant impact for VIP community contributing to the improvement in their quality of life providing safer, affordable, and reliable independent mobility in their local outdoor surroundings. It is expected that the smart and timely interactions of VIP with their surroundings through a direct live feed of interpretations of obstacles will significantly raise their level of confidence.

Expenditure (direct, indirect) for visual impairment across the globe in 2020 is estimated to be $2.8 Trillian (Adam, Lynne, & Cutler, 2020) posing a significant challenge in terms of expensive healthcare and preventing a large number of VIPs from getting the appropriate treatment. Technological advancement in this context may therefore be of noticeable advantage reducing the associated costs by delivering quality and affordable services (Burton et al., 2021). To this end, it is evident that the economic effectiveness of the proposed low-cost SOMAVIP is substantial, particularly for the concerned authorities (e.g., National Health Service in UK) that allocate large proportions of public funds for the local mobility of VIP and other direct and indirect costs to support the ODOMOVIP.

6. Conclusion and future directions

A deep transfer learning approach is proposed in this study utilising two object detection models (YOLOv5s and Mark R-CNN), trained for a customized dataset specific to ODOMOVIP, enabling an assistive technology for VIP. As per the RNIB report (Wilson, 2015) and other related literature identifying specific objects and obstacles as potential hazards to ODOMOVIP, a custom dataset from a variety of publicly available resources was constructed and made publicly available to the research community. The dataset was carefully annotated and prepared for the training and validation of the selected object detection models. Models were then trained for the annotated datasets comprising the list of objects (identified in RNIB reports (Wilson, 2015), (Aspirot, 2021) and other sources (Rnib, 2020)) and evaluated based on the test split consisting of random variations (e.g., rotations, translations, geometric features, background, image size, resolution) to measure the performance. ObDtM with high real-time object classification accuracy for purely unseen diverse dataset clearly demonstrated the robustness of the proposed approach as compared to existing object detection models. Statistical outcomes for the conducted experiments showed the development of reliable and efficient models (particularly with respect to the processing time) as required in the given context of ODOMOVIP. The ObDtM developed model along with the annotated dataset has been made publicly available and is accessible via (https://dx.https://doi.org/10.21227/sm6r-nb95) as described in the supplementary information. Although the ObDtM approach was developed for the objects of 8 different classes, the proposed model can be easily generalised by expanding its identification capability for other object classes with considerably moderate tuning for newly acquired information, thus enhancing its applicability domain. Additionally, the annotated custom dataset can be useful for the research community for continuous improvement with respect to model fine tuning, validation and even the development of new approaches for similar domains. Finally, an end-to-end smart framework (SOMAVIP) was proposed to enable the ODOMOVIP while utilising the developed object detection models along with other smart devices and digital services to transform the current VIP mobility aid into a fully autonomous data driven tool readily available for the next generation of responsive devices, IoT, and the emergence of smart city infrastructures. The proposed framework is expected to make significant contribution towards the improvement of VIP's QoL, including safer and independent ODM, psychological & mental well-being, and improved emotional and social aspects. Future work for the proposed approach is expected with the inclusion of other objects of different classes which can considered as potential hazards to ODMOBVIP (e.g., scaffolding, potholes from various parts over the globe) and embedding of obstacle detection devices such as LiDAR to generate a real time comprehensive commentary and perception of the surroundings.

7. Availability of data

The annotated dataset used in this article along with the custom-trained models is publicly available (https://dx.https://doi.org/10.21227/sm6r-nb95) as described in the supplementary information.

CRediT authorship contribution statement

Wasiq Khan: Conceptualization, Methodology, Software, Data curation, Visualization, Formal analysis, Writing – original draft, Project administration. **Abir Hussain:** Writing – original draft, Resources, Supervision. **Bilal Muhammad Khan:** Software, Validation, Writing – review & editing, Visualization. **Keeley Crockett:** Validation, Writing – review & editing.

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 is shared as described in Dataset section

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eswa.2023.120464.

References

Ackland, P., Resnikoff, S., & Bourne, R. (2017). World blindness and visual impairment: Despite many successes, the problem is growing. *Community Eye Health, 30*(100), 71–73.

Adam, G., Lynne, P., & Cutler, H. (2020). *The Global Economic Cost of Visual.* Access Economics Pty Limited. Retrieved 08 12, 2021, from.

Al-Fahoum, A., Al-Hmoud, H., & Al-Fraihat, A. (2013). A smart infrared microcontroller-based blind guidance system. *Act. Passive Electron. Compon*, 1–7.

Ali, M. (2017). Blind navigation system for visually impaired using windowing-based mean on microsoft kinect camera. *in Proc. 4th Int. Conf. Adv. Biomed. Eng. (ICABME)*, (pp. 1-4).

Aspirot, M. (2021). E-scooters to hit the streets again despite safety concerns. Retrieved August 15, 2021, from *Yahoo News, CBC.* https://ca.news.yahoo.com/e-scooters-hit-streets-again-223937053.html.

Bauer, Z., &,, et al. (2020). Enhancing perception for the visually impaired with deep learning techniques and low-cost wearable sensors. *Pattern Recognition Letters, 137*, 27–36. https://doi.org/10.1016/j.patrec.2019.03.008

Bauer., Z., & al., e.. (2020). Enhancing perception for the visually impaired with deep learning techniques and low-cost wearable sensors. *Pattern Recognition Letters, 137*, 27–36. https://doi.org/10.1016/j.patrec.2019.03.008

Burton, M., &,, et al. (2021). The Lancet Global Health Commission on Global Eye Health: Vision beyond 2020. *THE LANCET GLOBAL HEALTH COMMISSION, 09*(04), 459–551. https://doi.org/10.1016/S2214-109X(20)30488-5

Cao, Z., Liao, T., Song, W., Chen, Z., & Li, C. (2021). Detecting the shuttlecock for a badminton robot: A YOLO based approach. *Expert Systems with Applications, 164*. https://doi.org/10.1016/j.eswa.2020.11383

Capi, G., & Toda, H. (2012). Development of a new robotic system for assisting visually impaired people. *Int. J. Social Robot, 04*(01), 33–38.

Chen, Y., Liu, S., Shen, X., & Jia, J. (2020). DSGN: Deep Stereo Geometry Network for 3D Object Detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 12536-12545). IEEE.

Choi, S., Zhou, Q.-Y., Miller, S., & Koltun, V. (2016). *A Large Dataset of Object Scans.* arXiv. doi:arXiv:1602.02481.

Choi., S., Zhou., Q.-Y., Miller., S., & Koltun, V. (2016). *A Large Dataset of Object Scans.* arXiv. doi:arXiv:1602.02481.

DENG, Y., Luo, P., Loy, C. C., & Tang, X. (2014). Pedestrian Attribute Recognition At Far Distance. *Proceedings of the 22nd ACM international conference on Multimedia*, (pp. 789-792). doi:https://doi.org/10.1145/2647868.2654966.

DENG., Y., Luo., P., Loy., C. C., & Tang, X. (2014). Pedestrian Attribute Recognition At Far Distance. *Proceedings of the 22nd ACM international conference on Multimedia*, (pp. 789-792). doi:https://doi.org/10.1145/2647868.2654966.

Duarte, K., Cećılio, J., Silva, J. S., & Furtado, P. (2014). Information and Assisted Navigation System for Blind People. *Proceedings of the 8th International Conference on Sensing Technology.* Liverpool, UK.

Froneman, T., Heever, D. v., & Dellimore, K. (2017). Development of a wearable support system to aid the visually impaired in independent mobilization and navigation. *in Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, (pp. 783–786). Koria. doi:10.1109/EMBC.2017.8036941.

Giarre, C. L., &,, et al. (2019). An indoor and outdoor navigation system for visually impaired people. *IEEE Access, 07*, 170406–170418.

Giarre., C. L., & al., e.. (2019). An indoor and outdoor navigation system for visually impaired people. *IEEE Access, 07*, 170406–170418.

He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2017). Mask R-CNN. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, (pp. 2961-2969).

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *arXiv*. doi:10.48550/ARXIV.1512.03385.

Hoang, V., Nguyen, T., Le, T., &,, et al. (2017). Obstacle detection and warning system for visually impaired people based on electrode matrix and mobile Kinect. *Vietnam J Comput Sci, 04*(02), 71–83. https://doi.org/10.1007/s40595-016-0075-z

Howard, A. G., & et.al. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv.org*. Retrieved from https://arxiv.org/abs/1704.04861.

Hub, A., Hartter, T., & Ertl, T. (2006). Interactive localization and recognition of objects for the blind. *21st Annual International Technology and Persons with Disabilities Conference.*

Jaime Sanchez, C. O. (2011). Mobile audio assistance in bus' transportation for the blind. *International Journal on Disability and Human Development, 10*(04).

Jocher, G., Stoken, A., Borovec, J., Changyu, L., Hogan, A., Diaconu, L., et al. (2020). YOLOv5. *Retrieved from.* https://doi.org/10.5281/zenodo.4154370

Khan, W., & Kuru, K. (2021). A Framework for the Synergistic Integration of Fully Autonomous Ground Vehicles With Smart City. *IEEE Access, 09*, 923–948. https://doi.org/10.1109/ACCESS.2020.3046999

Khan., S., Nazir., S., & Khan, H. U.. (2021). Analysis of Navigation Assistants for Blind and Visually Impaired People: A Systematic Review. *IEEE Access, 09*, 26712–26734. https://doi.org/10.1109/ACCESS.2021.3052415

Khan., W., Hussain., A., Khan., B., Nawaz., R., & Baker, a. T. (2019). Novel Framework for Outdoor Mobility Assistance and Auditory Display for Visually Impaired People.

12th International Conference on Developments in eSystems Engineering (DeSE) (pp. 984-989). Kazan, Russia: IEEE. doi:10.1109/DeSE.2019.00183.

Krause., J., Stark., M., Deng., J., & Fei-Fei, L. (2013). 3D Object Representations for Fine-Grained Categorization. *4th IEEE Workshop on 3D Representation and Recognition, ICCV (3dRR-13).* Sydney, Australia.

Kumar., Y., & al., e.. (2010). RFID and GPS Integrated Navigation System for the Visually Impaired. *53rd International Midwest Symposium on Circuits and Systems (MWSCAS)*.

Kunta., V., C. Tuniki, & Sairam, U. (2020). Multi-Functional Blind Stick for Visually Impaired People. *5th International Conference on Communication and Electronics Systems (ICCES)* (pp. 895-899). India: IEEE. doi:doi: 10.1109/ICCES48766.2020.9137870.

Lee., Y. H., & Medioni, G. (2015). Wearable RGBD indoor navigation system for the blind. *in Proc. Eur. Conf. Comput. Vis*, (pp. 493–508).

Liao., C., Choe., P., Wu., T., Tong., Y., Dai., C., & Liu., Y. (2013). RFID-based road guiding cane system for the visually impaired. *in Proc. Int. Conf. CrossCultural Design* (pp. 86-93). Springer. doi:https://doi.org/10.1007/978-3-642-39143-9_10.

Lin., B.-S., Lee., C.-C., & P-Y, C.. (2017). Simple Smartphone-Based Guiding System for Visually Impaired People. *Sensors, 17*(06). https://doi.org/10.3390/s17061371

Lin., T., & et al. (2014). Microsoft COCO: Common Objects in Context. *ECCV, Part V, LNCS*, (pp. 740–755).

Lin., T.-Y., Dollar., P., Girshick., R., He., K., Hariharan., B., & Belongie, S.. (2017). Feature Pyramid Networks for Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2117–2125).

Lin., Y., Wang., K., Yi., W., & Lian, S.. (2019). Deep Learning Based Wearable Assistive System for Visually Impaired People. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Nakajima, M., & S. h.. (2013). New indoor navigation system for visually impaired people using visible light communication. *EURASIP J. Wireless Commun. Netw, 01* (01), 37.

Mattoccia, M. P. (2016). A wearable mobility aid for the visually impaired based on embedded 3D vision and deep learning. In *IEEE Symposium on Computers and Communication (ISCC)* (pp. 208–213). https://doi.org/10.1109/ISCC.2016.7543741

AtikurRahman, M.d., & M. s.. (2021). IoT Enabled Automated Object Recognition for the Visually Impaired. *Computer Methods and Programs in Biomedicine*. *Update, 01*(01). https://doi.org/10.1016/j.cmpbup.2021.100015

Nada., A. A., Fakhr., M. A., & Seddik, A. F. (July, 2015). Assistive infrared sensor based smart stick for blind people. *in Proc. Sci. Inf. Conf. (SAI), pp. 1149–1154.*

Ni., D., Song., A., Tian., L., Xu., X., & Chen, D.. (2015). A walking assistant robotic system for the visually impaired based on computer vision and tactile. *International Journal of Social Robotics, 07*. https://doi.org/10.1007/s12369-015-0313-z

Nick, Z. G. (2018). *An Introduction to Evaluation Metrics for Object Detection.* Retrieved 08 02, 2021, from https://blog.zenggyu.com/en/post/2018-12-16/an-introduction-to-evaluation-metrics-for-object-detection/.

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering, 22*(10), 1345–1359.

Parikh., N., Shah., I., & Vahora, S. (2018). Android Smartphone Based Visual Object Recognition for Visually Impaired Using Deep Learning. *International Conference on Communication and Signal Processing (ICCSP)* (pp. 0420-0425). IEEE. doi:10.1109/ICCSP.2018.85.

Pinedo, M., & Villanueva., F., Santofimia., M., & Lopez, J.. (2011). Multimodal positioning support for ambient intelligence. In *Proceedings of the 5th International Symposium on Ubiquitous Computing and Ambient Intelligence* (pp. 1–8).

Redmon, J., & Farhadi, A. (2018). *YOLOv3: An incremental improvement.* arXiv: 1804.02767. Retrieved from https://pjreddie.com/media/files/papers/YOLOv3.pdf.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, , pp. 779–788. Las Vegas, NV, USA.

Ren, S., He, K., Girshick, R., & Sum, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *In Proceedings of the Neural Information Processing System (NIPS)*, (pp. 1–9). Canada.

Rist, C. B., Emmerichs, D., Enzweiler, M., & Gavrila, D. M. (2022). Semantic Scene Completion Using Local Deep Implicit Functions on LiDAR Data. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(10). https://doi.org/10.1109/TPAMI.2021.3095302

Rnib. (2020). Highlighting the e-scooter challenge to safety on our streets. from *UK: RNIB. Retrieved, 11 12*, 2021 https://www.rnib.org.uk/news/campaigning/e-scooter-challenge-streets-safety.

Russakovsky., O., & al., e.. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV), 115*, 211–252. https://doi.org/10.1007/s11263-015-0816-y

S., M., & P., M. (2015). 3D Glasses as Mobility Aid for Visually Impaired People. In B. M. Agapito L. (Ed.), *Computer Vision - ECCV 2014, Workshops, Lecture Notes in Computer Science. 8927.* Springer, Cham.

Sanchez, M. S. (2008). Orientation and mobility in external spaces for blind apprentices using mobile devices. *Mag. Ann. Metrop. Univ, 08*, 47–66.

Sanchez., J., & Saenz, a. M. (2008). *Orientation and mobility in external spaces for blind apprentices using mobile devices.*

Shao., L., Han., J., Kohli., P., & Zhang, Z. (2014). RGB-D Sensor-Based Computer Vision Assistive Technology for Visually Impaired Persons. In *Computer Vision and Machine Learning with RGB-D Sensors. Advances in Computer Vision and Pattern Recognition.* Springer, Cham. doi:https://doi.org/10.1007/978-3-319-08651-4_9.

Simoes., W. C., & Lucena, V. F. (2016). Hybrid indoor navigation as sistant for visually impaired people based on fusion of proximity method and pattern recognition algorithm,'. *in Proc. IEEE 6th Int. Conf. Consum. Electron.* Berlin (ICCE-Berlin).

Skulimowski., P., Owczarek., M., Radecki., A., Bujacz., M., Rzeszotarski., D., & Strumillo, P.. (2019). Interactive sonification of U-depth images in a navigation aid for the visually impaired. *J. Multimodal User Interface, 13*(03), 219–230.

Srivastava, S., Divekar, A. V., Anilkumar, C., Naik, I., Kulkarni, V., & Pattabiraman, V. (2021). Comparative analysis of deep learning image detection algorithms. *Journal of Big Data, 08*. https://doi.org/10.1186/s40537-021-00434-w

Sylvie Treuillet, E. R. (2010). Outdoor/indoor vision-based localization for blind pedestrian navigation assistance. *International Journal of Image and Graphics, 10*(04), 481–496.

Tan, L., Huangfu, T., Wu, L., & Chen, W. (2021). Comparison of RetinaNet, SSD, and YOLO v3 for real-time pill identification. *BMC Medical Informatics and Decision Making.* https://doi.org/10.1186/s12911-021-01691-8

W., L., & al., e. (2016). SSD: Single Shot MultiBox Detector. *In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science. 9905.* Springer, Cham. doi:https://doi.org/10.1007/978-3-319-46448-0_2.

Wachaja.A., Agarwal.P., Adame.M.R., Möller.K., & Burgard.W. (Sep. 2014, pp. 13–14). A navigation aid for blind people with walking disabilities. *in Proc. IROS Workshop Rehabil. Assistive Robot.*

WHO. (2021). *Blindness and vision impairment.* World Health Organisation. Retrieved 12 07, 2021, from https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment.

Wilson, M. (2015). *Who Put That There: The barrier to blind and partially sighted people getting out and about.* UK: Royal National Institute of Blind People. Retrieved March 12, 2018, from https://www.rnib.org.uk/sites/default/files/Who%20put%20that%20there%21%20Report%20February%202015.pdf.

Yang., C., Huang., Q., Li., Z., Liu., K., & Hu., F.. (2017). Big Data and cloud computing: Innovation opportunities and challenges. *International Journal of Digital Earth, 10* (01), 13–53. https://doi.org/10.1080/17538947.2016.1239771

Zegarra, J., & Farcy, R. (2012). GPS and inertial measurement unit (IMU) as a navigation system for the visually impaired. *in Proc. Int. Conf. Comput. Handicapped Persons.*

Zhao, R., Pang, M., & Zhang, Y. (2018). Robust shape extraction for automatically segmenting raw LiDAR data of outdoor scenes. *International Journal of Remote Sensing, 39*(23). https://doi.org/10.1080/01431161.2018.1508914

Zöllner, M., Huber, S., Jetter, H., & Reiterer, H. (2011). NAVI—A proof-ofconcept of a mobile navigational aid for visually impaired based on the microsoft Kinect. In *Human-Computer Interaction – INTERACT* (Vol. 6949, pp. 584–587). Springer, Berlin, Heidelberg. doi:https://doi.org/10.1007/978-3-642-23768-3_88.